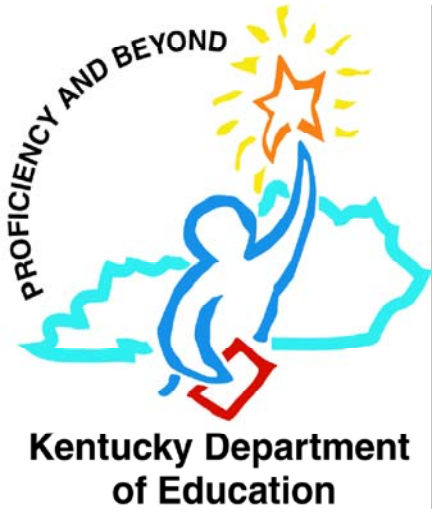
Kentucky Core Content Tests

2000
Technical Report



Based on the Analysis of Data from the 1999-2000 School Year

Produced by the Kentucky Department of Education
January 2002

Acknowledgements

Our thanks to the following contributors:

Scott Trimble
Associate Commissioner, Office of Assessment and Accountability,
Kentucky Department of Education

Bill Insko
Director, Division of Assessment Implementation, Kentucky Department of Education

Robert Wetter
Senior Research Consultant, Kentucky Department of Education

Linda Frazer
Director, Division of Research and Validation, Kentucky Department of Education

Michael Burdge
University of Kentucky

Mary Kanalakis
Senior Program Director, CTB/McGraw-Hill

Tammy Bullock
Program Manager, CTB/McGraw-Hill

Deborah Molin
Program Office Coordinator, CTB/McGraw-Hill

Ross Green
Chief Research Psychologist, CTB/McGraw-Hill

Hillary Michaels
Senior Research Scientist, CTB/McGraw-Hill

James Carlson
Chief Research Scientist, CTB/McGraw-Hill

Michelle Boyer
Research Associate, CTB/McGraw-Hill

John Hayter
Research Associate, CTB/McGraw-Hill

Erica Connelly
Research Associate, CTB/McGraw-Hill

Marilyn Gehrman
Senior Project Manager, State Assessment Programs, Data Recognition Corporation

Jim McMann
Project Manager, State Assessment Programs, Data Recognition Corporation

Bob Kampa
Director, Performance Assessment, Data Recognition Corporation

David Payne
Project Manager, Writing Portfolio, Data Recognition Corporation

Joanne Jensen
Director of Test Development, WestEd

Kentucky Core Content Tests
2000 Technical Report

Table of Contents

Appendices:

Chapter 1
Introduction

Introduction

In 1989 the Kentucky Supreme Court deemed the entire system of public elementary and secondary education in Kentucky unconstitutional. The Court also directed the Kentucky General Assembly to create and enact into law a new system of education that was not only constitutional but also based upon efficiency as defined by adequacy and equity. The result was House Bill 940, the Kentucky Education Reform Act (KERA), which was enacted to provide an "adequate education for all students" as mandated by the courts. One of the most comprehensive, statewide restructuring efforts ever attempted in the United States, the reform called for systemic change in finance, governance, curriculum and assessment. With regard to Kentucky's assessment system, KERA required the establishment of learning goals and identified procedures for defining and assessing the new goals. The following bullets provide an overview of the events that lead to KERA:

- November 1985 – The Council for Better Education, a nonprofit corporation formed by 66 school districts, seven boards of education, and 22 public school children sued the state of Kentucky for not providing an efficient system of education.

- October 1988 – Franklin County Circuit Court Judge Ray Corns found for the plaintiffs.

- February 1989 – Through his own actions, Governor Wallace Wilkinson issued an executive order creating a twelve-member Council on School Performance Standards. The Council was charged with determining what all students should know and be able to do and how learning should be assessed.

- June 1989 – the Kentucky Supreme Court directed the General Assembly to recreate and reestablish a "new efficient system of common schools" that complied with the Kentucky Constitution. The Court defined an efficient system of common schools as an organization that provides a "free and adequate education to all students throughout the state regardless of geographical location or local fiscal resources."

- September 1989 – the Council on School Performance Standards produced the report *Preparing Kentucky Youth for the Next Century: What Students Should Know and Be Able To Do and How Learning Should Be Assessed* and presented it to the Curriculum Committee of the Legislative Task Force charged with creating Kentucky's new system. Six broad learning goals for all students were recommended with particular emphasis on what they should be able to do. In addition, the Council recommended that the state launch a major effort to assess student performance beyond what can be measured by paper-and-pencil tests. It also was recommended that the state initiate long-range development efforts that support school reform in implementing the new learning goals.

- In 1990, the Council's recommendations were incorporated into House Bill 940, the Kentucky Education Reform Act, as a first step in redefining the school curriculum and providing what the courts required as an adequate education for all students.

- April 11, 1990 – House Bill 940 was signed by Governor Wallace Wilkinson and became law on July 13, 1990. With KERA, the General Assembly established the framework for a major revision of Kentucky's educational system. KERA required the establishment of learning goals for the educational system, provided a procedure by which those goals would be defined and assessed, and created a series of rewards and assistance to be associated with the performance of schools on those assessments.

The six learning goals established by KERA for schools within the Commonwealth are presented in the following table.

Table 1-1
Kentucky School Goals

Goal 1	Expect a high level of achievement of all students.
Goal 2	Develop student's abilities in six cognitive areas.
Goal 3	Increase school attendance rates.
Goal 4	Reduce dropout and retention rates.
Goal 5	Reduce physical and mental health barriers to learning.
Goal 6	Increase the proportion of students who make a successful transition to work, postsecondary education, and the military.

Through a two-year period of public input and review, 75 valued outcomes or performance goals were produced. The Kentucky Board of Education (KBE) approved these in December of 1991. Concerns arose about the measurability of learner goals three and four (see Table 1-1), and complaints were made about the obscurity of the wording of the valued outcomes. These concerns led to the revision and reduction of the valued outcomes to 57 in number. These were presented to the Kentucky Board of Education on May 3-4, 1994. Since that time, they have been known as the Academic Expectations. In addition to the Learning Goals and Academic Expectations, in 1992 the Kentucky Instructional Results Information System (KIRIS) was developed to measure progress toward the goals, primarily the expectations reflected in the first two goals of the act, and the non-cognitive goals outlined in goals three, four and six.

In 1998, House Bill 53 made adjustments to Kentucky's assessment and accountability programs, creating a new system call the Commonwealth Accountability and Testing System, or CATS. More specifically, an important part of this legislation directed the Kentucky Board of Education to redesign the assessment and accountability system. Through a broad and collaborative process involving educators and citizens of Kentucky, many changes were made in this new system first administered in the spring of 1999. The changes were made in order to improve the reliability and validity of the test, reduce testing time and make the system fairer and easier to understand. Those changes include, but are not limited to:

- Distributing the test components for the high school from primarily the junior year to across three grade levels;

- Reducing the contents of the Writing Portfolio in each accountability year;
- Limiting the student to answers on the open response to the space provided—one 8 ½″ x 11″ sheet;
- Including multiple-choice questions on the Kentucky Core Content Tests and weighting them 33% of the score, and the open response at 67% of the Kentucky Core Content Test component of the Commonwealth Accountability Testing System;
- Giving schools incremental credit for Novice and Apprentice growth in reading, math, science and social studies; and,
- Reducing the testing window from 3 weeks to 2 weeks.

House Bill 53 shaped Kentucky's assessment and accountability system through several provisions that outline general features of a system of testing and biennial school accountability, leaving many details of implementation to various committees that were enacted by the bill. For example, the School Curriculum, Assessment, and Accountability Council (SCAAC) was created by House Bill 53 to study, review, and make recommendations concerning Kentucky's system of setting academic standards, assessing learning, holding schools accountable for learning, and assisting schools to improve their performance. The council advises the Kentucky Board of Education (KBE) and the Legislative Research Commission (LRC) on issues related to the development and communication of the Academic Expectations and Core Content for Assessment, and the development and implementation of the statewide assessment and accountability program, including the distribution of rewards and imposition of sanctions. SCAAC is composed of 17 voting members appointed by the Governor. The appointments are made to assure broad geographical representation and representation of elementary, middle, and secondary school levels, as well as equal representation of the two sexes, inasmuch as possible, and to assure that appointments reflect the minority racial composition of the Commonwealth.

House Bill 53 also required the Legislative Research Commission to appoint a National Technical Advisory Panel on Assessment and Accountability (NTAPAA), which must be composed of no fewer than three professionals with a variety of expertise in education testing and measurement. The panel advises LRC, and upon approval of the Director of the Commission, the Kentucky Board of Education and the Department of Education.

In addition to the above legislation, state law also requires KBE to set policy and promulgate regulations to implement both the assessment and accountability systems. The following are a few of the more important regulations promulgated by KBE:

703 KAR 5:010 Writing portfolio procedures.

703 KAR 5:020 The formula for determining school performance classifications and school rewards.

703 KAR 5:040 Statewide Assessment and Accountability Program; relating accountability index to school classification.

703 KAR 5:050 Statewide Assessment and Accountability Program; school building appeal of performance judgments.

703 KAR 5:070 Procedures for the inclusion of special populations in the state-required assessment and accountability programs.

703 KAR 5:080 Administration Code for Kentucky's Educational Assessment Program.

703 KAR 5:120 Assistance for schools; guidelines for scholastic audit.

703 KAR 5:130 School district accountability.

703 KAR 5:140 Requirements for school and district report cards.

Interim Accountability Model for 2000

Kentucky's accountability system is a high-stakes system with rewards and sanctions attached to results. The over-riding goal of the CATS is for all schools in Kentucky to reach Proficiency as defined by the Kentucky Board of Education. The accountability system provides the mechanism for measuring this goal and thus provides feedback to schools on how they are progressing toward the long-term goal set by the Kentucky Board of Education. Schools achieving rewards status in CATS receive money from the state to be used for school purposes. According to a recent Attorney General's Opinion, "for school purposes" *does* include the use of reward money for teacher bonuses.

For the accountability cycle ending in 2000, over 20 millions dollars was distributed to schools achieving rewards status. Schools falling short of their goal at the end of a particular cycle, by regulation (703 KAR 5:120, see above), receive a Scholastic Audit, receive the assistance of a Highly Skilled Educator, and are eligible to receive state funds to be targeted toward improvement. The Scholastic Audits performed by state, regional and local district personnel are thorough and provide audited schools information on over 80 indicators related to school success. While Kentucky's accountability system is based upon measuring continued improvement toward a long-term goal, and thus has built in monitoring to ensure real and enduring improvement, the Scholastic Audits contribute to this monitoring by focusing on those schools that need assistance the most. It should be noted that a school selected for a Scholastic Audit may or may not be a school targeted for Title I funds.

The Long-Term Accountability model (703 KAR 5:020), to take effect in 2002, is a growth model with schools serving as their own baseline. All students and thus all schools are expected to demonstrate improvement within the system. Because of the major changes in the system imposed by House Bill 52 in 1998, comparisons between KIRIS (pre-1998) and CATS (post-1998) are not appropriate. Words like 'gain', 'growth', 'improvement', or 'decline' are not appropriate ways to describe the difference between 1997 and 1998 scores on KIRIS and the 1999 and 2000 Kentucky Core Content Tests (KCCT) results of the CATS. Because of this lack of ability to compare the two tests, neither the old (pre-1998) nor the new Long-Term Accountability models are appropriate for determining rewards and assistance in the year 2000. To solve this problem, the National Technical Panel for Assessment and Accountability advised the State Board of Education to use a regression-based model using 1997 and 1998 KIRIS data to predict 1999 and 2000 performance. The State Board selected the model after months of discussion and upon the recommendation of NTAPAA. The panel, which has advised the Board on various technical aspects of developing the new testing system, characterized the model as

statistically sound and offering Kentucky the ability to compare results during the transition biennium (1999 and 2000) between the state's old testing system and the new testing system. More information about this can be found in Chapter 13 (Interim Accountability) of this report and on the Kentucky State Department of Education website.

Measures and Indicators

Both academic content-based and non-academic measures were used in KIRIS, and both are still used in CATS. These measures include custom, criterion-referenced tests in reading, mathematics, science, social studies, arts and humanities, practical living/vocational studies and writing. Non-academic measures include attendance rate, retention rate, dropout rate and transition to adulthood. (Note that transition to adulthood data is collected in the fall of each year via a short survey completed by school personnel. Measures include the number of graduates planning to enter college, the military, or an alternative vocation.) The above multiple measures were selected to provide as complete a snapshot of schools as possible and to communicate to schools the importance of each measure and indicator in terms of resources and instructional programs. All measures used in CATS are state mandated; as such there is no role for local school/district assessment information in the accountability system.

As stated earlier, the long-term goal for every school in the state is Proficiency as defined by the Kentucky Board of Education. This goal of Proficiency translates into a school accountability index value of 100 (i.e., the goal for the state is for each school to achieve an accountability index of 100 by 2014). Each of the measures/indicators mentioned above are combined into a composite to obtain a school's accountability index. Through an initial standard setting process in the summer of 1992 and an independent verification of the standards in 1995, the goal of Proficiency, and thus an index value of 100, was directly related to a description of how students have to perform to achieve Proficiency and is thus directly related to the indicators that are part of the assessment system.

It should be noted that for the biennium ending in 2002, a lengthy standard setting process (see Chapter 8 on Standard Setting), was undertaken to redefine the goal of Proficiency, and thus an index value of 100, and this time was directly related to specific grade level and content area descriptions of *what* students have to know, and *how* students have to perform, to achieve Proficiency. As such, the construct of "good school" or "improving school" for 2002 is defined by specific content and performance standards and is directly related to the indicators that are part of the assessment system.

The following table summarizes the grades and content areas tested by the Kentucky Core Content Test (KCCT), including the number of open-response and multiple-choice questions asked on each of six forms of the KCCT (12 forms each for arts and humanities and practical living/vocational studies). Because there are six forms of the test and forms generally do not overlap, this means that for accountability purposes there are 36 open-response items and 144 multiple-choice items administered per grade level/content area for reading, mathematics, science and social studies. For arts and humanities and practical living/vocational studies there are 24 open-response items and 96 multiple-choice items administered per grade level/content

area because there are 12 non-overlapping forms of the test. In addition, students at grades 4, 7 and 12 select and respond to one of two on-demand writing prompts offered during the test.

1999-2000 ASSESSMENT COMPONENTS									
Grade	**Kentucky Core Content Test**							**Portfolio**	
	Rdg	**Math**	**Sci**	**Soc St**	**Wrtg**	**A&H**	**PL/VS**	**Wrtg**	**Alt***
4	**6 OR* 24 MC**		**6 OR 24 MC**		**X***			**X**	**X**
5		**6 OR 24 MC**		**6 OR 24 MC**		**2 OR 8 MC**	**2 OR 8 MC**		
7	**6 OR 24 MC**		**6 OR 24 MC**		**X**			**X**	
8		**6 OR 24 MC**		**6 OR 24 MC**		**2 OR 8 MC**	**2 OR 8 MC**		**X**
10	**6 OR 24 MC**						**2 OR 8 MC**		
11		**6 OR 24 MC**	**6 OR 24 MC**	**6 OR 24 MC**		**2 OR 8 MC**			
12					**X**			**X**	**X**

*** OR denotes Open Response, MC denotes Multiple Choice; "X" denotes that On-Demand Writing (or the Writing Portfolio) was administered; "Alt" denotes participation in the Alternative Portfolio program.**

All testing is completed in the spring of each year, including the administration of a norm-referenced test in grades 3 (end of primary), 6 and 9. Beginning in 2002 the results of the norm-referenced test will contribute to the calculation of a schools accountability index. In addition to the KCCT, non-academic components used in accountability include attendance rate and retention rate for all schools, in middle and high schools - dropout rate, and in high school, a measure of transition to adulthood.

Kentucky's assessment program offers accommodated or modified assessments for students who qualify. The accommodation/modification must be stipulated in the student's Individual Education Plan (IEP) or 504 and must have been used with the student throughout the school year. For example, if a student's IEP allows a scribe during regular instruction, the student will be allowed to have a scribe for the statewide assessment. Other accommodations or modifications, when consistent with the normal on-going delivery of instruction, may include:

- Reading text in English
- Paraphrasing directions for tasks in English
- Oral word-for-word translation of text
- Administering assessments in small groups
- Use of foreign language dictionaries
- Use of word processor or typewriter
- Use of grammar or spell-checker.

In addition to the above accommodations or modifications, Kentucky has a two-year exemption for students whose primary language is not English. More specifically, Limited English Proficient (LEP) students must have been in an English-speaking school for two full years

preceding the year of the assessment before participating in the assessment with or without accommodations or modifications. Because this policy is not in alignment with federal regulation (i.e., Title I and IDEA), Kentucky applied for and has been granted a one-year exemption while the state develops policies for serving and assessing LEP students. Depending upon the current reauthorization, the state plans on allowing only a one-year exemption for LEP students prior to participating in the statewide assessment.

Students who cannot participate in the regular assessment, even with accommodations, are required to submit an alternate portfolio. These students usually have profound cognitive disabilities and the alternate portfolio is the only way they can participate in the assessment and accountability systems. With few exceptions, all students in Kentucky must participate in the regular assessment or the alternate portfolio. Students can receive a medical exemption if certain criteria are met (e.g., the stated medical condition *cannot* be the student's disability) and a physician determines that the student cannot physically take the test or that participation would be harmful to the child. Foreign exchange students are also exempt from the statewide assessment. All together, less than one percent of students statewide are exempted each year from Kentucky's assessment program.

Kentucky's Accountability Index

The long-term goal for every school in the state is Proficiency as defined by the Kentucky Board of Education. The goal of Proficiency translates into a school accountability index value of 100. More specifically, the goal for the state is for each school to achieve an accountability index of at least 100 by 2014. In the Long-Term Accountability Model referenced above, intermediate targets that will eventually take a school to the goal of 100 are set biennially, or every two years starting in 2002. As such, there are seven biennia or accountability cycles between 2002 and 2014 (i.e., 2002, 2004, 2006, 2008, 2010, 2012 and 2014). The major characteristics of the accountability model is that it involves (a) an index, (b) comparisons or a measure of growth between successive groups, (c) criteria that are applicable to the whole school and (d) differential weighting of indicators. While the Interim Accountability Model for 2000 is based upon a regression model for establishing an expectation for each school's performance, the same characteristics apply except (b) -- a measure of growth.

With respect to the Interim Accountability Model, the previously discussed indicators are combined to create an accountability index that is unique to each school. The progression of how this happens begins with simple number-correct raw scores and ends with an accountability index that summarizes a school's progress toward the state's goal of Proficiency. To state this progression in one sentence, raw scores give rise to scale scores, scale scores have been related to Novice, Apprentice, Proficient and Distinguished (NAPD) performance levels (via standard setting and cut-scores), NAPD's get weighted numerically and combined within each content area, and finally, the content areas are weighted and combined to form a school's accountability index. The following 4 steps describe this process in more detail.

Step 1 - Raw Scores Give Rise to Scale Scores

As previously stated in the Measures and Indicators section, there are multiple forms of the test for each grade level and content area assessed and the forms generally do not overlap. To compensate for small differences in difficulty among forms, and to bring all forms of a test for a grade level and content area onto the same scale, Item Response Theory is used. As such the underlying scale for the KCCT is not number-correct raw score, but rather a scale score scale that ranges from approximately 325 to 800 with 500 being the middle of the scale. (Note that in 1997 and 1998, a theta scale was used that ranged from approximately –3 to +3.)

Step 2 - Scale Scores Have Been Related to Performance Levels

It can be argued that the heart and soul of both KIRIS and CATS is the four performance levels used to describe the quality of student work. The four levels, from lowest to highest, are Novice, Apprentice, Proficient and Distinguished or NAPD. During standard setting, these four performance levels were related to, or mapped onto, the range of scale scores for each grade level and content area test. In addition, beginning in 1999, the first two levels of performance in reading, mathematics, science and social studies have each been subdivided into three levels (Novice non-performance, Novice medium, Novice high, Apprentice low, Apprentice medium and Apprentice high) to better represent student performance.

Step 3 - NAPD's Get Weighted Numerically and Combined

Students taking a test in a particular content area are assigned to one of the above eight performance levels. This is the official "score" that gets reported for the student. For example, a forth grade student might receive an Apprentice in reading and a Proficient in science. For reporting in the aggregate and for accountability purposes only, the following conversion table is used for transforming NAPD's into a numerical scale that ranges from 0 to 140:

Performance Level	Weight
• Novice Non-performance	0
• Novice Medium	13
• Novice High	26
• Apprentice Low	40
• Apprentice Medium	60
• Apprentice High	80
• Proficient	100
• Distinguished	140

For example, if the following distribution (or percentages) were obtained by fourth graders administered the reading test in a particular school, the calculations would be:

Performance Level	Weight	Percentage	Calculation
• Novice Non-performance	0	5%	0 X .05
• Novice Medium	13	10%	13 X .10
• Novice High	26	15%	26 X .15
• Apprentice Low	40	20%	40 X .20
• Apprentice Medium	60	25%	60 X .25
• Apprentice High	80	15%	80 X .15
• Proficient	100	8%	100 X .08
• Distinguished	140	2%	140 X .02
• Total of Sum		100%	51.0

As demonstrated in the above table, the weights for the NAPD's are multiplied by the percentages (or rather the proportions) of students at each performance level and then simply summed across the performance levels. The resulting content area index for fourth grade reading in this school is 51.0. The same procedure is used for calculating the "academic" index for each content area. Note the direct connection between the performance levels and a content area or academic index. If every fourth grade student in the school had scored Proficient (i.e., the state goal) on the reading test, the school reading index would be 100 (or at the state goal). As seen in the next step, this connection is maintained all the way through to a school's weighted accountability index.

Step 4 - Content Areas Get Weighted and Combined

Once an academic index has been calculated for all content area tests administered within a school, the school's accountability index for a particular year can then be determined. The weights used to calculate a school's accountability index vary slightly depending upon whether the school is an elementary, middle or high school. The following formula reflects the weighting of components at the *high school* level.

Given the following definition of terms in the formula:

RD = Reading	AH = Arts & Humanities
MA = Mathematics	PL = PL/VS
SC = Science	WR = Writing
SS = Social Studies	NA = Non-academic

To calculate the index for a given year:

Accountability Index = (RD*.15) + (MA*.15) + (SC*.15) + (SS *.15) +
(WR*.15) + (AH*.075) + (PL *.075) + (NA*.10).

The weights used for calculating an Accountability Index sum to one. Note that the above formula, or weighted composite, for the Accountability Index is for one year only. However, the above Accountability Index calculations have to be performed for both years of the baseline and both years of the subsequent target years. For example, the Interim Accountability baseline index is the arithmetic mean of the Accountability Index for 1997 and for 1998, i.e., (1997 Index + 1998 Index)/2. In the same way, the target index for the Interim Accountability Cycle ending in 2000 is the arithmetic mean of the Accountability Index for 1999 and for 2000, or (1999 Index + 2000 Index)/2. These two indices (i.e., the baseline and the target) are the values used in the regression analysis for the Interim Accountability Cycle.

Other important considerations regarding Kentucky's Accountability Index include:

- Because many schools in Kentucky are small, two years of data are combined to form both the baseline and the growth indices. Combining two years of data addresses some of the stability issues related to estimating achievement for small schools. The Interim Accountability Model was used to evaluate all regular schools (and students within alternative programs) regardless of school size. As such, there is *not* a special review process for small schools.
- Results from non-standard administrations of the assessment (accommodated or modified testing) are included in accountability calculations the same way as results from standard administrations of the tests.
- While K-2 schools do not participate in the assessment program which starts in grade 3 (end of primary), these schools can receive reward money if the regular or accountable school the K-2 school feeds into qualifies for rewards. (It should be noted that there were only 19 K-2 or K-3 schools in Kentucky during the 1999/2000 school year. Of those, seven K-3 schools actually had waivers in place to have their accountability scores included with the "receiving" school.)
- The four non-academic components (i.e., attendance, retention, dropout and successful transition to adult life) are not computed on the 0 to 140 scale. Rather, these components are each put onto a 0 to 100 scale. More specifically, the values for attendance and successful transition to adult life are the actual percentages reported, whereas the values entered into calculations for retention and dropout are 100 minus the actual percentage calculated. Because of the minimal weighting attributed to non-cognitive measures, the impact of this on a school's overall, weighted accountability index is slight.
- For Title I, an index is created for each district based only upon the schools within the district that receive Title I funds. This index is evaluated for purposes of federal reporting.

As a final note, results from the Alternate Portfolio, Kentucky's means of assessing the instruction provided to students with significant disabilities, are scored using the same performance levels as the content area tests (i.e., NAPD). An Alternate Portfolio is submitted only once at the elementary level, once at the middle school level, and once at the high school level. *At each of these levels, a student's performance level (N, A, P or D) weight contributes to all content areas*. For example, if an Alternate Portfolio student receives a Proficient, for calculation purposes, it is as if the student received a Proficient (weight of 100) in all content areas of the assessment at the

grade level. In this way, Alternate Portfolio students contribute the same amount to accountability as any other regular education student, although that contribution happens within one calendar year and not across several years (e.g., fourth and fifth grade or seventh and eighth grade). The main justification for this is the importance of including all students in assessment and accountability. Similarly, the scores for students who receive accommodations or modifications are treated the same as students who received no accommodations or modifications. In Kentucky, the inclusion of all students is weighed more heavily, i.e., is more important in terms of consequential validity, than the small challenge to construct validity that may result when alternate and accommodated student scores are included with all other student scores.

Purpose of This Technical Report

The purpose of this technical report is to provide information about the technical characteristics of the 2000 Interim Accountability Cycle of CATS. A secondary purpose is to track the changes that have occurred to the system during the time span covered by this report. The time period under consideration is the Interim Accountability Cycle, which spanned the four school years 1997 through 2000. While some parts of this report are accessible to everyone, its intended audience is experts in psychometrics and educational research. This report is best understood with a working knowledge of measurement concepts such as reliability and validity, and statistical concepts such as correlation and central tendency. For some chapters, the reader is presumed to have basic familiarity with advanced topics in measurement and statistics such as item response theory and factor analysis.

This Technical Report provides extensive detail about the development and operation of the Kentucky Core Content Tests. The traditional concerns with a program are often labeled reliability and validity. The empirical reliability and validity of the assessments are reported explicitly in this document. While the reliability chapter (Chapter 11) is relatively straightforward, the validity chapter (Chapter 12) addresses the validity of the program as derived from the sum of its parts. That is, the steps in creating the program and putting it into operation are all aspects of validity. The validity of any assessment stems from the steps taken in planning it, the processes of developing the content of the tests, the processes of consulting with stakeholders, the processes of communicating about the test to users, the processes of scoring and reporting, and the processes of data analysis. Each is an inherent part of validity. The uses made of the test results as established by the Kentucky Board of Education are also aspects of validity. In short, while there is a specific chapter devoted to validity, including many references to validity types of studies, this document provides much, but not all of the evidence needed to assess the validity of the program.

In reading this technical report, it is critical to remember that the testing program does not exist in a vacuum; it is not just a test. It is one part of a complex network intended to help schools focus their energies on dramatic improvement in student learning. CATS is an integrated program of testing, accountability, and curricular and instructional support, coupled with wide-sweeping changes in school finance, governance, and organization. It can only be evaluated properly within this full context.

Chapter 2
Test Specifications

Introduction

There are two primary reasons for test specifications. First, they help ensure that the assessment is measuring what it is intended to measure. Second, test specifications help ensure that across the four years of an accountability cycle, the sample of tasks upon which school success is judged are equivalent.

Three sets of specifications were used to develop the 1999-2000 Commonwealth Accountability and Testing System (CATS) Kentucky Core Content Test (KCCT)—the Academic Expectations, the Core Content for Assessment, and the Test Blueprint. The Academic Expectations characterize what students are to achieve and are tied to Kentucky's six learning goals. The learning goals broadly define the achievement expectations for all students. The Core Content for Assessment provides greater definition and specification of the content that will be included in the KCCT while the Test Blueprint indicates the relative emphasis of the content outlined in the Core Content for Assessment. Each of these documents is considered below. In addition to the aforementioned specifications, Kentucky's curriculum frameworks provide benchmarks and further information about content, concepts, and context for questions on the KCCT.

Learner Goals and Academic Expectations

Like KIRIS, the KCCT assesses four of the six learner goals—goals 1, 2, 5 and 6. Table 2-1 lists the four goals that are the basis of the assessment. By statute, goals 3 and 4 are not assessed. Goals 5 and 6 are addressed through the writing portfolio.

Table 2-1
Kentucky's Four Learner Goals That Are Measured by the KCCT

1. Students are able to use basic communication and math skills for purposes and situations they will encounter throughout their lives.
2. Students shall develop their abilities to apply core concepts and principles from mathematics, the sciences, the arts, the humanities, practical living studies, and vocational studies to what they will encounter throughout their lives.
3. Students shall develop their abilities to think and solve problems in a variety of situations they will encounter in life.
4. Students shall develop their abilities to connect and integrate experiences and new knowledge from all subject matter fields with what they have previously learned, and build on past learning experiences to acquire new information through media sources.

These learning goals have been further defined by 57 academic expectations. These statements provide global statements about the expectations of what students should know and be able to do. (The 1995-96 Technical Report provides the history of the development of the academic expectations.)

Core Content for Assessment

The Core Content for Assessment, developed by the Kentucky Department of Education, provides information to educators about the focus of the assessment. Recognizing the difference between information that should be taught but not necessarily assessed (because of the sensitive nature of a topic or the inability of a paper and pencil assessment to adequately reflect student learning), the Core Content for Assessment provides guidelines for the development and selection of the assessment items. This document defines the assessable content at three levels—subdomain, section, and specific content statements for all grades and subjects. For example for grade 7 science, the content code 2.3.2 corresponds to the subdomain of "Earth and Space Science", the section entitled "Earth in the Solar System", and the specific content statement, "Most objects in the solar system are in regular and predictable motion. Those motions explain such phenomena as the day, the year, phases of the moon, and eclipses". All subject areas follow the same convention except for reading that does not specify content at the section level.

A revised Core Content for Assessment was published in September 1999. The Kentucky Department of Education Division of Curriculum Development was responsible for the revision. Working with teams of teachers for each grade and subject represented in the KCCT, the Core Content for Assessment was reviewed and updated to reflect the committees' current thinking as to the appropriate content for assessment beginning with the Spring 2000 KCCT.

The revised Core Content was first used by the Content Advisory Committees (CAC's) for item selection for the Spring 2000 KCCT. The document was presented to the CAC's at the item selection meeting held in September 1999. Kentucky Department of Education Curriculum Development staff presented the revised document to the individual committees and prepared a crosswalk document which summarized the relationship between the previous version of the Core Content and the September 1999 revision.

The changes in the Core Content did impact item selection. Some items that had previously matched the Core Content were found to no longer match the revised document. This mismatch was found for both matrix and pretest items. Consequently, matrix items were repeated rather than selecting items that did not match the revised Core Content. Pretest items were revised, if possible, to provide a match to the revised content. (The degree of allowable edits was limited. Major revisions were not possible because significant edits would have necessitated a review by the Bias Committee prior to their inclusion.) The exception was for the linking form. Beginning with the Spring 2000 assessment, an entire test form was identified to provide a statistical link to the 1999 assessment. The matrix items for the linking form were repeated in the Spring 2000 assessment even if they did not match the revised Core Content.

Test Blueprint

The Test Blueprint is a public document designed to communicate the structure and contents of the Kentucky Core Content Test to classroom teachers, administrators, school councils, and other interested persons.[1] It also is designed to provide guidelines for item development and selection.

The same teacher committees who approved the final draft of the Core Content recommended how the test content should be represented in the KCCT. In contrast to the Test Blueprint for the 1999 KCCT that provided specifications for test content at both the subdomain and section level, the Test Blueprint (version 3.0) provides proportions only for the subdomain level for each content area. Like the 1999 document, overall percentages are reported. Separate breakdowns are not provided for multiple-choice and open-response items. Tables 2-2 through 2-7 provide the test blueprint for all grades and subjects assessed by the Spring 2000 KCCT.

Table 2-2
Test Blueprint (Version 3.0)
For Reading

Subdomain	Grade 4	Grade 7	Grade 10
Literary Reading	50%	40%	30%
Informational Reading	25%	25%	30%
Persuasive Reading	10%	15%	20%
Practical/Workplace Reading	15%	20%	20%
Total	**100%**	**100%**	**100%**

Table 2-3
Test Blueprint (Version 3.0)
For Mathematics

Subdomain	Grade 5	Grade 8	Grade 11
Number/Computation	45%	35%	20%
Geometry/Measurement	25%	25%	30%
Probability/Statistics	15%	15%	15%
Algebraic Ideas	15%	25%	35%
Total	**100%**	**100%**	**100%**

[1] Taken from The Kentucky Core Content Test Blueprint (version 3.0).

Table 2-4
Test Blueprint (Version 3.0)
For Science

Subdomain	Grade 4	Grade 7	Grade 11
Physical Science	33%	30%	35%
Earth/Space Science	33%	35%	30%
Life Science	33%	35%	35%
Total	**100%**	**100%**	**100%**

Table 2-5
Test Blueprint (Version 3.0)
For Social Studies

Subdomain	Grade 5	Grade 8	Grade 11
Government and Civics	25%	30%	20%
Culture and Society	12%	15%	10%
Economics	10%	10%	15%
Geography	25%	15%	20%
Historical Perspective	28%	30%	35%
Total	**100%**	**100%**	**100%**

Table 2-6
Test Blueprint (Version 3.0)
For Arts and Humanities

Subdomain	Grade 5	Grade 8	Grade 11
Music	30%	25%	25%
Dance	20%	20%	20%
Drama/Theatre	20%	20%	20%
Visual Arts	30%	25%	25%
Literature	-----	10%*	10%
Total	**100%**	**100%**	**100%**

* This percentage applies only to multiple-choice questions. No open-response items are to be developed for literature for grade 8. The open-response questions that would have been allocated to literature are to be equally divided between music and visual arts.

Table 2-7
Test Blueprint (Version 3.0)
For Practical Living/Vocational Studies

Subdomain	Grade 5	Grade 8	Grade 10
Health	40%	35%	35%
Physical Education	25%	20%	10%
Consumerism	15%	15%	15%
Vocational Studies	20%	30%	40%
Total	**100%**	**100%**	**100%**

Actual Test Content Coverage

Consistent with the Test Blueprint specifications, Tables 2-8 through 2-13 provide breakdowns at the subdomain level for each grade for the live test items that appeared in the Spring 2000 KCCT. These values are based only on the primary content code, although an individual item can receive up to three content codes. These tables include the total number of multiple-choice and open-response items for the 1999-2000 KCCT. The values are reported separately for linking and matrix items. Items that are repeated are counted each time they appear in the test. The percentage of items by subdomain as specified by the Test Blueprint is provided for comparison.

Table 2-8
Distribution of Items Across Subdomains for
Grade 4 Multiple-Choice and Open-Response

Content Area	Subdomain	Linking		Matrix		Total	% of Total of All Items in Subject	% in Blueprint
		MC	OR	MC	OR			
Reading	Literary Reading	12	3	60	15	90	50%	50%
Reading	Informational Reading	8	2	36	9	55	31%	25%
Reading	Persuasive Reading	0	0	12	3	15	8%	10%
Reading	Practical/Workplace Reading	4	1	12	3	20	11%	15%
Science	Physical Science	5	1	40	8	54	30%	33.3%
Science	Earth/Space Science	8	2	36	8	54	30%	33.3%

Table 2-9
Distribution of Items Across Subdomains for
Grade 5 Multiple-Choice and Open-Response

Content Area	Subdomain	Linking		Matrix		Total	% of all Items In Subject	% in Blueprint
		MC	OR	MC	OR			
Arts and Humanities	Music	6	1	23	6	36	30%	30%
Arts and Humanities	Dance	3	0	16	5	24	20%	20%
Arts and Humanities	Drama/Theatre	2	1	17	4	24	20%	20%
Arts and Humanities	Visual Arts	5	2	24	5	36	30%	30%
Mathematics	Number/ Computation	10	2	49	10	71	39%	45%
Mathematics	Geometry/ Measurement	7	2	33	8	50	28%	25%
Mathematics	Probability/ Statistics	5	1	20	8	34	19%	15%
Mathematics	Algebraic Ideas	2	1	18	4	25	14%	15%
Practical Living/ Vocational Studies	Health	5	2	38	11	56	46%	40%
Practical Living/ Vocational Studies	Physical Education	7	1	20	3	31	26%	25%
Practical Living/ Vocational Studies	Consumerism	1	1	9	2	13	11%	15%
Practical Living/ Vocational Studies	Vocational Studies	3	0	13	4	20	17%	20%
Social Studies	Government and Civics	7	1	17	7	32	18%	25%
Social Studies	Culture and Society	1	1	21	2	25	14%	12%
Social Studies	Economics	2	2	19	6	29	16%	10%
Social Studies	Geography	5	1	30	6	42	23%	25%
Social Studies	Historical Perspective	9	1	33	9	52	29%	28%

Table 2-10
Distribution of Items Across Subdomains for
Grade 7 Multiple-Choice and Open-Response

Content Area	Subdomain	Linking		Matrix		Total	% of Total of All Items In Subject	% in Blueprint
		MC	OR	MC	OR			
Reading	Literary Reading	9	2	44	11	66	36%	40%
Reading	Informational Reading	7	2	36	9	54	30%	25%
Reading	Persuasive Reading	4	1	20	5	30	17%	15%
Reading	Practical/Workplace Reading	4	1	20	5	30	17%	20%
Science	Physical Science	9	1	40	12	62	34%	30%
Science	Earth/Space Science	9	3	38	8	58	32%	35%
Science	Life Science	6	2	42	10	60	33%	35%

Table 2-11
Distribution of Items Across Subdomains for
Grade 8 Multiple-Choice and Open-Response

Content Area	Subdomain	Linking		Matrix		Total	% of all Items	% in Blueprint
		MC	OR	MC	OR			
Arts and Humanities	Music	5	2	19	6	32	27%	25%
Arts and Humanities	Dance	4	0	13	4	21	18%	20%
Arts and Humanities	Drama/Theatre	2	1	21	4	28	23%	20%
Arts and Humanities	Visual Arts	4	1	24	6	35	29%	25%
Arts and Humanities	Literary	1	0	3	0	4	3%	10%
Mathematics	Number/ Computation	6	1	39	7	53	29%	35%
Mathematics	Geometry/ Measurement	8	3	30	6	47	26%	25%
Mathematics	Probability/ Statistics	5	1	22	11	39	22%	15%
Mathematics	Algebraic Ideas	5	1	29	6	41	23%	25%
Practical Living/ Vocational Studies	Health	7	1	30	8	46	38%	35%
Practical Living/ Vocational Studies	Physical Education	2	1	14	3	20	17%	20%
Practical Living/ Vocational Studies	Consumerism	2	0	12	4	18	15%	15%
Practical Living/ Vocational Studies	Vocational Studies	5	2	24	5	36	30%	30%
Social Studies	Government and Civics	3	2	33	3	41	23%	30%
Social Studies	Culture and Society	3	1	15	3	22	12%	15%
Social Studies	Economics	4	1	20	7	32	18%	10%
Social Studies	Geography	7	2	17	5	31	17%	15%
Social Studies	Historical Perspective	7	0	37	12	56	31%	30%

Table 2-12
Distribution of Items Across Subdomains for
Grade 10 Multiple-Choice and Open-Response

Content Area	Subdomain	Linking		Matrix		Total	% of Total of All Items In Subject	% in Blueprint
		MC	OR	MC	OR			
Practical Living/ Vocational Studies	Health	8	2	27	6	43	36%	35%
Practical Living/ Vocational Studies	Physical Education	0	1	8	2	11	9%	10%
Practical Living/ Vocational Studies	Consumerism	1	0	16	4	21	17%	15%
Practical Living/ Vocational Studies	Vocational Studies	7	1	29	8	45	38%	40%
Reading	Literary Reading	4	1	40	10	55	31%	30%
Reading	Informational Reading	8	2	24	6	40	22%	30%
Reading	Persuasive Reading	4	1	24	6	35	19%	20%
Reading	Practical/Workplace Reading	8	2	32	8	50	27%	20%

Table 2-13
Distribution of Items Across Subdomains for
Grade 11 Multiple-Choice and Open-Response

Content Area	Subdomain	Linking		Matrix		Total	% of all Items In Subject Area	% in Blueprint
		MC	OR	MC	OR			
Arts and Humanities	Music	4	0	22	5	31	26%	25%
Arts and Humanities	Dance	1	1	12	3	17	14%	20%
Arts and Humanities	Drama/Theatre	4	0	14	4	22	18%	20%
Arts and Humanities	Visual Arts	5	2	20	6	33	28%	25%
Arts and Humanities	Literature	2	1	12	2	17	13%	10%
Mathematics	Number/ Computation	6	1	37	6	50	28%	20%
Mathematics	Geometry/ Measurement	6	3	32	11	52	29%	30%
Mathematics	Probability/ Statistics	7	1	20	3	31	17%	15%
Mathematics	Algebraic Ideas	5	1	31	10	47	26%	35%
Science	Physical Science	11	2	43	11	67	37%	35%
Science	Earth/Space Science	3	1	25	5	34	19%	30%
Science	Life Science	10	3	52	14	79	44%	35%
Social Studies	Government and Civics	4	2	22	6	34	19%	20%
Social Studies	Culture and Society	1	1	9	7	18	10%	10%
Social Studies	Economics	5	1	21	4	31	17%	15%
Social Studies	Geography	5	0	23	4	32	18%	20%
Social Studies	Historical Perspective	9	2	45	9	65	36%	35%

The Spring 2000 KCCT was the first assessment based on the revised Core Content (Version 3.0) and Test Blueprint. The expectation was that by 2002 the distribution of items across each subdomain would be within 5% of the blueprint recommendations, and that there might be slightly greater discrepancies in 2000 due to the lack of sufficient items based on the revised Core Content. For 56 of the 74 subdomains (75%) the distribution was within 5% of the blueprint recommendations. For 17 subdomains, the discrepancies fell between 6 and 9%.

It should be noted that the previous Core Content for Grade 8 Arts and Humanities included a list of specific literary works and authors upon which items could be based. The revised Core Content removed this list. All test items that referred to a specific literary work and required prior knowledge of the work were removed from the 2000 KCCT. This resulted in a total of only 3% of the items reflecting literature yielding a discrepancy of 7% below the recommendations of the Test Blueprint for that subdomain. Science was the content area that experienced the most significant change in its Core Content, which is reflected in the content coverage match to the blueprint. The greatest discrepancy was 11% in grade 11 Science for the subdomain Earth and Space Science.

In order to increase the match to the Test Blueprint for the 2001 KCCT, WestEd consultants reviewed the existing test coverage to identify the subdomains in need of development. The CAC development of pretest items was guided by this review.

Number of Test Questions Per Student

The change in the basic test design that occurred with the move from Kentucky Instructional Results and Information Systems to the Commonwealth Accountability Testing System was maintained in the Spring 2000 assessment. Six base forms were developed for reading, mathematics, science and social studies with two alternate versions labeled A and B. These alternate forms provided for the administration of embedded pretest items. The number of items that appeared for each subject in each form is reported in Table 2-14. For arts and humanities and practical living/vocational studies, the number of items that students complete is less than for the other subject areas. Consequently, the A/B versions of each test form present unique matrix and pretest items for these two subject areas in order to provide adequate content coverage.

Table 2-14
Number of Items Taken by Student by Subject Area

	Matrix		Pretest	
	MC	OR	MC	OR
Reading	24	6	4	1
Mathematics	24	6	4	1
Science	24	6	4	1
Social Studies	24	6	4	1
Arts and Humanities	8	2	4	1
Practical Living/Vocational Studies	8	2	4	1

The ideal test design would yield 144 unique multiple-choice and 36 unique open-response items for reading, mathematics, science, and social studies. The corresponding values for arts and humanities and practical living/vocational studies are 96 and 24. As indicated previously, the change in the Core Content for Assessment impacted the item selection process. Items that had previously appeared in the KCCT had to be dropped from consideration for the 2000 KCCT because of the lack of match to the revised Core Content. Table 2-15 provides the number of unique live items for each grade and subject for the Spring 2000 assessment.

Table 2-15
Number of Unique Items By Item Type

Grade	Subject	Question Type	Number of Unique Items
04	Reading	Multiple-Choice	136
04	Reading	Open-Response	34
04	Science	Multiple-Choice	137
04	Science	Open-Response	33
05	Arts and Humanities	Multiple-Choice	96
05	Arts and Humanities	Open-Response	24
05	Mathematics	Multiple-Choice	144
05	Mathematics	Open-Response	36
05	Practical Living/Vocational Studies	Multiple-Choice	96
05	Practical Living/Vocational Studies	Open-Response	24
05	Social Studies	Multiple-Choice	144
05	Social Studies	Open-Response	36
07	Reading	Multiple-Choice	144
07	Reading	Open-Response	36
07	Science	Multiple-Choice	116
07	Science	Open-Response	35
08	Arts and Humanities	Multiple-Choice	95
08	Arts and Humanities	Open-Response	24
08	Mathematics	Multiple-Choice	144
08	Mathematics	Open-Response	34
08	Practical Living/Vocational Studies	Multiple-Choice	96
08	Practical Living/Vocational Studies	Open-Response	24
08	Social Studies	Multiple-Choice	145*
08	Social Studies	Open-Response	36
10	Practical Living/Vocational Studies	Multiple-Choice	96
10	Practical Living/Vocational Studies	Open-Response	24
10	Reading	Multiple-Choice	144
10	Reading	Open-Response	35
11	Arts and Humanities	Multiple-Choice	96
11	Arts and Humanities	Open-Response	24
11	Mathematics	Multiple-Choice	136
11	Mathematics	Open-Response	33
11	Science	Multiple-Choice	116
11	Science	Open-Response	36
11	Social Studies	Multiple-Choice	141
11	Social Studies	Open-Response	33

* An edit was made to a matrix item in Form A of the test but not Form B. Consequently, a new item number had to be assigned to the edited item.

Summary

Kentucky's Academic Expectations, Core for Assessment, and the Test Blueprint all served to provide specifications for the selection of items to be included on the KCCT. Although the Academic Expectations did not change, the Core Content for Assessment and the Test Blueprints did change. Although the changes in Core Content were viewed as minor in terms of curricular change, they did have an impact on the item selection process. Items that had previously matched the Core Content had to be dropped from the 2000 assessment. This resulted in the repeating of items across test forms and a reduction in the overall number of items in the item pool as shown in Table 2-15. The subject area most impacted by this change was science. The change affected not only the "live" items but also newly developed pretest items. Where possible, pretest items were revised, but significant revisions were not possible because the schedule for test production did not allow for another review by the Bias Review Committee.

As Content Advisory Committee members worked with the Core Content for Assessment, they identified content statements that required clarification and/or further specification. It is recommended that the Department develop a guide that is consistent with the interpretations used by the Content Advisory Committees so that all teachers are informed of the scope of content to be assessed.

In the future, it is suggested that any changes to the Core Content for Assessment be introduced at the item development phase rather than item selection. This would ensure that newly developed pretest items would match the new document and allow the test development process to more immediately reflect any changes.

Chapter 3
Test Development Process

Introduction

In contrast to the 1998–1999 assessment year, the 1999–2000 assessment cycle provided the opportunity to develop and select items for the KCCT over the course of the year. This allowed the development of pretest items in the spring, a full editorial review by WestEd staff, an evaluation by KDE curriculum staff of the edited items, and Bias Committee Review of edited items prior to Content Advisory Committee (CAC) selection meetings. The test development process outlined in this chapter provides a description of the procedures followed during the 1999–2000 assessment year from item development through test form production.

Content Advisory Committees

The item development and selection process is based on the work of the Content Advisory Committees. The members of these committees are appointed by the Kentucky Department of Education. Each grade and subject included in the KCCT is represented by a Content Advisory Committee. These committees of ten involve classroom teachers, school administrators, and university personnel. These representatives are drawn from throughout the service center regions to ensure geographic balance. Similarly, efforts are made to provide ethnic representation for each grade and subject committee. Although ten members are selected for the committees, not all committee members attend on a regular basis. The average number of participants per meeting was seven. For several committees there were as few as five CAC members in attendance. The lack of full committee participation places an added burden on those committee members who are present, particularly during item development because the total number of items to be developed by a given committee does not change based on the number of committee members attending a meeting.

The Test Development Sequence

Because the development cycle for the Spring 2000 assessment could be distributed throughout the year, the process began with the development of pretest items in May of 1999. The sequence of steps is outlined below.

Table 3-1

Major Steps in the Development Process, May 1999 through February 2000

1.	Comparison of content coverage for Spring 1999 KCCT to test blueprint to identify test development needs.
2.	Reading CACs submit passages for pre-test (April).
3.	Bias Review of potential pretest passages (May).
4.	CAC development of pretest items (May).
5.	CAC development of preliminary scoring guides for pretest open-response items (May).
6.	Content and editorial review of pretest items by WestEd editorial staff (May through August).
7.	Initial research for copyright permission for pretest passages and graphics/images (May through August).
8.	Development of pretest graphics (May through August).
9.	Submission of pretest items to KDE for content and editorial review, including evaluation of content codes (August).
10.	Preparation of pretest open-response item summaries based on preliminary scoring information provided by Data Recognition Corporation (August).
11.	Preparation of item statistics for CAC review (August/September).
12.	Identification of linking form by CTB (September).
13.	Bias Committee review of pretest items (September).
14.	CAC review of 1999 KCCT data to determine which items do/do not meet statistical guidelines for inclusion in the KCCT (September).
15.	CAC evaluation of proposed items with respect to match to Core Content (September).
16.	Review and revision of scoring guides for selected open-response items (September).
17.	Review and selection of pretest items to meet Test Blueprint needs (September).
18.	Review of previously drafted pretest scoring guides (September).
19.	Assemble test forms (October through January).
20.	Review test forms for balance and cueing effects (October through January).
21.	Develop camera-ready final copy (November through February).

1. **Comparison of Content Coverage for Spring 1999 KCCT to Test Blueprint to Identify Test Development Needs.** Prior to the CAC item development meetings, WestEd consultants reviewed the existing test coverage to determine where the blueprint was met and where items needed to be developed to meet the blueprint by 2002. This step was particularly important this year because of the change in the Core Content and Test Blueprint. This activity identified the subdomains and specific content statements in need of development.

2. **Reading CACs Submit Pretest Passages.** Members of the Reading Content Advisory Committees were asked to submit reading passages for pretest item development to WestEd. The committee members were provided the following guidelines for the number of words per passage by grade level (Table 3-2). They were also provided with forms for submitting information regarding the source of the passages to ensure that the necessary information

was available to pursue copyright permission. Once the passages were received, they were sorted by grade and prepared for review by the Bias Committee.

Table 3-2
Desired Reading Passage Length in Number of Words

	Short	Medium	Long
Grade 4	500	600-900	1000-1100
Grade 7	500-600	700-1000	1100-1500
Grade 10	600	700-1300	2000

3. **Bias Committee Review of Pretest Passages.** The passages submitted to WestEd by CAC members were then submitted to the Bias Review Committee for consideration. The role of this committee was to ensure that the content of the passages was fair and equitable for all students, and that the passages did not contain material that could be considered stereotyped racially, ethnically, regionally, economically, or gender-stereotyped or biased toward any group. Committee members received training prior to passage review based on the *Guidelines for Handling Sensitive Issues in Kentucky's State Assessment Development.* The purpose of this review was to identify any passages that the Bias Committee did not find acceptable prior to the item development meeting, so that CAC members would only have acceptable passages to choose from when developing new items.

4. **CAC Development of Pretest Items.** Following training on the development of assessment items, the individual Content Advisory Committees met separately to develop new pretest items. Committee members were given writing assignments based on identified areas of need according to the Test Blueprint. As part of the item development process, the full committee reviewed all items in their draft form. A critical aspect of this development phase was the review and match of the item content to the grade-appropriate Core Content statements. WestEd consultants maintained records of committee recommendations for item changes.

5. **CAC Development of Preliminary Scoring Guides for Pretest Open-Response Items.** A key step in the development of open-response items is a determination of how students are expected to respond to a given item. Committee members drafted sample student responses, and these responses were compared to the wording of the question to ensure there was a match between what was being asked in the question and what was expected in a student's response.

6. **Content and Editorial Review by WestEd Editorial Staff.** Following the CAC meetings, WestEd consultants reviewed each item for content accuracy. Items were then submitted to WestEd's editorial staff who reviewed each item for adherence to WestEd's test development guidelines. All proposed edits were reviewed with WestEd's content consultants. During the editorial review, WestEd conducts a readability analysis of all the proposed pretest reading passages. This analysis includes a word count and a review of five different readability indices (Dale-Chall, Flesch, FOG, Powers, and SMOG). Each of these

readability indices provides a different index as to the overall readability of a given passage. Because there is no one index that is viewed as the best indicator of passage difficulty, all five indices are reviewed. If a passage is found to be above grade level by all five formulas, the passage is dropped from consideration. Primary consideration is given to the Flesch, FOG and SMOG indices.

7. **Initial Research for Copyright Permission for Pretest Passages and Graphics/Images.** The CACs recommended passages and images (e.g., photographs, drawings, maps,) for pretest consideration, and each was researched to determine if permission could be obtained for use in an assessment. WestEd developed forms for passage and graphics submission to ensure that the necessary information was available to pursue copyright permission.

8. **Development of Pretest Graphics.** Graphics have become a key component of the KCCT. Whereas some graphics were obtained directly from permission agencies (e.g., photographs or paintings for arts and humanities, maps for social studies), others had to be created by WestEd's desktop publishers. Pretest item reviews were not completed until the necessary graphics were reviewed with the item**.**

9. **Submission of Pretest Items to KDE for Content and Editorial Review, Including Evaluation of Content Codes.** Although KDE staff from the Curriculum Development Division participated in the item development process, it was determined that their review of the pretest items prior to selection was critical to ensure the match to Core Content and Kentucky's instructional emphasis. KDE curriculum consultants reviewed items for content accuracy and match to the grade-appropriate Core Content. These consultants noted any questions or concerns about the item's content or coding directly on the items. This feedback was provided to the CACs during the item selection meetings.

10. **Preparation of Pretest Open-Response Item Summaries Based on Preliminary Scoring Information Provided by Data Recognition Corporation (DRC).** Before open-response pretest items could be considered for inclusion as a matrix item, each was reviewed for quality. WestEd staff developed summaries of each pretest item drawing on scored student work and evaluations from DRC scoring directors. DRC scoring directors were asked to provide feedback on ease of scoring, clarity of the scoring guide, clarity of the item as judged by student responses. In addition, they were asked to recommend changes to both the wording of the item and the wording of the scoring guide. WestEd staff then produced a written summary to be used by the Content Advisory Committee members. The items were judged based on the clarity of student responses, range of student responses, and ease of scoring. These summaries, combined with actual student work reflecting the range of student performance, formed the basis for the evaluation of pretest open-response items and their corresponding scoring guides.

11. **Preparation of Item Statistics for CAC Review.** In order to determine what items should be considered for inclusion in the Spring 2000 assessment, item-level analyses were conducted by CTB. The results of these item-level analyses were combined with printouts of the items as they appeared in the test books. These materials formed the basis of item selection.

12. **Identification of Linking Form by CTB.** In contrast to previous years in which the CACs identified individual items for the purposes of providing cross-year linking, beginning with the Spring 2000 KCCT, an entire test form was identified as linking.

13. **Bias Committee Review of Pretest Items.** Following KDE review of the pretest items, all of the pretest items (including accompanying graphics) were submitted to the Bias Review Committee for consideration. The role of this committee was to ensure that the content of the items was fair and equitable for all students, and that the items did not contain material that could be considered stereotyped racially, ethnically, regionally, economically, or gender-stereotyped or biased toward any group. Committee members received training prior to item review based on the *Guidelines for Handling Sensitive Issues in Kentucky's State Assessment Development.*

    The committee evaluated each item to determine if it was acceptable as submitted, acceptable with revision, or rejected. The judgments of this committee were viewed as advisory. If after consultation with KDE staff an item was deemed acceptable with revision, an item that had been rejected by the Bias Review Committee could be included in the assessment.

14. **CAC Review of 1999 KCCT Data to Determine Which Items Do/Do Not Meet Statistical Guidelines for Inclusion in the KCCT.** The statistical performance of an item in the 1999 KCCT was central to the item selection process. Item-level data were reviewed to determine which items from the previous year's assessment could be considered for inclusion in the 2000 KCCT. Table 3-3 provides the statistical guidelines used by the Content Advisory Committees to review, select, and where necessary, edit multiple-choice items. Table 3-4 provides the corresponding guidelines used for open-response items. Prior to item selection, the Content Advisory Committees received training on the interpretation of item statistics, the statistical guidelines to be used in item selection, the test blueprint, and test design.

Table 3-3
Multiple-Choice Item Statistics Used in the Item Selection Process

Item mean
Percent of students selecting each response option
Biserial correlation for the correct answer
Biserial correlation for incorrect response options

Table 3-4
Open-Response Item Statistics Used in the Item Selection Process

Item mean
Item standard deviation
Percent of students scoring at 0, 1, 2, 3 or 4 level for a given item
Item-level to total score correlation
Number of blank responses
Number of nonblank responses

WestEd was provided general statistical guidelines by the Kentucky Department of Education to use in item selection. The guidelines for multiple choice items called for p-values ranging from 0.30-0.80, point biserial correlations for the correct responses greater than .15 and point biserial correlations for incorrect response options less than 0.0. For the open-response items, the statistical guidelines were not as specific. When evaluating open-response items, the emphasis was placed on selecting items that did not have a significant percentage of blanks or scores of zero. Blanks and zeroes were viewed as indicators of students' inability to access the item. The Department of Education requested that students be provided an entry point into each open-response item whenever possible. Following this consideration, items were selected to reflect a range of difficulty.

Not all items were found to achieve the desired guidelines, particularly in the areas of middle school and high school math and high school science, where a substantial number of multiple-choice items were found to be more difficult than desired. Table 3-5 lists the number of multiple-choice items with p-values outside the desired range. In contrast, Grade 4 reading and science, Grades 7 and 10 reading had 20 or more items that were easier than the recommended guideline. For reading, four multiple-choice items and one open-response item must be selected for each passage. In some cases, an item or two for a given passage was found to be easier than desired, but these items were included in an effort to meet the test blueprint rather than drop the entire passage from consideration. For Grade 4 science, the changes in Core Content required that items not be included because of their failure to match the revised document. This led to the selection of a number of items that had p-values greater than .80, but their inclusion was necessary to provide the best possible match to the test blueprint.

Table 3-5
Number of Multiple-Choice Matrix Items
with P-Values Outside the Desired Guidelines

Grade	Subject	p-value < 0 .3	p-value > 0.8
4	Reading	0	38
4	Science	6	30
5	Arts and Humanities	1	10
5	Mathematics	9	13
5	Practical Living/Vocational Studies	1	10
5	Social Studies	0	12
7	Reading	2	30
7	Science	2	9
8	Arts and Humanities	1	10
8	Mathematics	19	2
8	Practical Living/Vocational Studies	0	15
8	Social Studies	1	19
10	Practical Living/Vocational Studies	0	8
10	Reading	5	24
11	Science	20	9
11	Arts and Humanities	4	7
11	Mathematics	40	5
11	Social Studies	2	13

For all items where the item statistics (p-value and point biserials) fell outside the desired ranges, the Content Advisory Committees were asked to scrutinize these items specifically for match to Core Content and grade-level appropriateness. If the items were deemed appropriate, they were retained in the matrix.

Where item statistics provided support for changes, multiple-choice items were edited to provide clarification to improve item functioning. Some items had p-values that were in the acceptable range and acceptable point biserial correlations for the correct answer, while the point biserial correlation for one of the incorrect response options was too high thus indicating that some of the better performing students were drawn to that option. If an item was needed in order to provide content coverage (match to the test blueprint), an incorrect response option was revised to make it "less attractive" to students. Such changes were made only when a rationale could be provided for the change. The correct response options were not edited, but the order of response options may have changed. Item stems were edited to include changes in emphasis (add bold-faced text or all caps) or to correct minor errors in the item. If an item required a more substantial change (more than one incorrect response option or the correct answer), it was pretested again.

For existing open-response items, format changes were allowed. As with the multiple-choice items, changes in emphasis were made. Where it was believed that separating an item into component parts (e.g., parts a,b,c) would result in improved student performance, the item was divided into parts. Some of the items that had been pretested and were selected for the matrix were edited such that students were asked to do less than had been asked for previously (e.g., students may have been asked to provide four reasons to support a

conclusion on the pretest but asked for only three reasons when the item was used for matrix purposes). For all items (multiple choice and open response) the intent of the item never changed, nor were students asked to provide more information than was asked for previously.

16. **CAC Evaluation of Proposed Items with Respect to Match to Core Content.** Because of the critical nature of the match of every item selected for inclusion in the KCCT with the grade-appropriate Core Content, committee members reviewed the content codes and academic expectations assigned to every item. Individual items could receive up to three separate content codes/academic expectations. These codes were confirmed for each item selected.

17. **Review and Revision of Scoring Guides for Selected Open-Response Items.** Once open-response items were selected and edited as needed, the Content Advisory Committees reviewed the scoring guides to ensure that the expectations of student performance called for in the scoring guides were appropriate for the grade level and the item as worded. Revisions were made as appropriate. No changes were made to the scoring guides for linking items to ensure the same scoring standard was applied to these items across years.

18. **Review and Selection of Pretest Items to Meet Test Blueprint Needs.** Once the matrix items were selected, pretest items were chosen. The selection of pretest items was based on the areas of need as determined by the Test Blueprint. The Content Advisory Committees selected pretest items based on overall item quality, match to Core Content, and coverage need as determined by the Test Blueprint. Because of the changes in the Core Content for Assessment and the Test Blueprint from the time of the pretest development meeting to the introduction of the revised documents, some pretest items had to be revised in order to provide the match to the revised Core Content. Because pretest items had already been reviewed by the Bias Committee, the degree of allowable edits was limited. With the revisions in the Core Content, some subdomains were still not adequately represented in the pretest item pool.

19. **Review of Previously Drafted Pretest Scoring Guides.** Because pretest items were open to revision, it was critical to have the committees review the scoring guides once the items had been finalized. Committee members were asked to provide scoring information that would help the scorers in establishing Kentucky's scoring standard. WestEd consultants recorded and incorporated committee recommendations for changes in the pretest scoring guides.

20. **Assemble Test Forms.** Following all item selections, WestEd test development staff assembled the test forms. The balance of content coverage was the primary concern when constructing test forms. Items were assigned across test forms to provide a distribution of the related content across the forms. Following content coverage, items were assigned to balance item difficulty across forms. The visual complexity of the forms was addressed by attempting to balance the number of graphics across forms. Where necessary, response options were reordered so that no clear pattern emerged in the correct answer choices.

Item assignments also were affected by the designation of Form 1 for visually-impaired students and Form 2 for hearing-impaired students. The visually-impaired form required the selection of simple graphics and ones in which a black and white contrast was all that was necessary to interpret the item. For the hearing-impaired form, items that required students to be able to hear sounds or tones were not included. Arts and Humanities was the content area that was affected most significantly by the restrictions for developing visually-impaired and hearing-impaired forms. Form 3 was designated as the linking form.

21. **Review Test Forms for Balance and Cueing Effects.** Once the forms were assembled, each was reviewed to determine if there was an appropriate mixture of core content and academic expectations. In addition, items both within and across subject areas were reviewed for possible cueing of correct answers. Items were reassigned to forms as needed.

22. **Develop Camera-Ready Final Copy.** Following a series of editorial and proofing checks, WestEd sent camera-ready copy to Data Recognition Corporation (DRC) for printing. DRC also provided an editorial review of all test forms.

The Role of Core Content in the Item and Test Development Process

Ensuring the validity of the assessment is one of the primary responsibilities of the test development contractor. Throughout the item development and selection and the test form production process, content validity is of primary concern. The Core Content for Assessment is the document that serves as the guide to content validity.

In both the item development and item selection process, the Content Advisory Committee members were trained specifically on the importance of content validity to the test development process. Committee members were told that the match to Core Content was the most important factor for consideration in item development and selection. Committee members were admonished that every item developed or selected must reflect a **clear** match to the specified Core Content statement. To emphasize this point, committee members were asked to evaluate each item's match to Core Content by considering whether the proposed item is appropriate for assessment for any student in Kentucky who has been adequately taught the Core Content. As items were developed, the content code(s) assigned to each item were reviewed by the committee, not just the individual or individuals who wrote the item. Throughout WestEd's content and editorial review, the match to Core Content was evaluated. The review of all newly developed pretest items by Kentucky Department of Education staff specifically included an evaluation of the match to Core Content. Finally, as all items were selected for inclusion on the KCCT, all content codes were reviewed, revised as needed, and approved by the Content Advisory Committees. As per the direction from the Department, the Content Advisory Committees had the final approval of all coding related to Core Content.

Just as the Core Content guides the item development and selection process, the consideration of content plays an important role in form development. Form development requires a balance of both content coverage and item difficulty. As items were selected for inclusion on particular

forms, every effort was made to balance the content coverage across forms to ensure their comparability with respect to the Core Content being assessed. This provided for the best sampling of the content to be assessed across forms.

Development of Scoring Guides

WestEd and DRC shared responsibility for the development of scorer training materials. WestEd revised the scoring rubrics for the matrix open-response items based on the items' performance in the previous year and the recommendations of the CACs. The initial drafts of pretest scoring guides were developed in conjunction with the CACs and were further developed by WestEd test development staff. All scoring guides were forwarded to DRC prior to the receipt of student work. DRC scoring directors reviewed the scoring guides for clarity prior to reading student responses.

WestEd staff worked collaboratively with DRC scoring directors to identify anchor, training and qualifying papers to be used in the training of scoring staff. These papers were based on early returns received and processed by DRC prior to the summer 2000 scoring sessions. Where necessary, the scoring guides were edited to provide a clear match of the anchors to the language of the scoring guide. Multiple anchors were identified for each score point. Where there was more than one way to earn a given score point, an anchor was identified for each score option. Following the identification of anchors, training papers were selected. These papers served to both reinforce the anchors and provide scorers exposure to responses that were more problematic in that they did not clearly match the language of the scoring guide. Qualifying papers were used to evaluate whether scorers were appropriately interpreting the scoring guide and accurately scoring sample papers before scoring live student work.

Summary

The test development and item selection process involved the Content Advisory and Bias Review Committees and staff from the Kentucky Department of Education. Both the CACs and Bias Review committees involved broad representation from the different regions of Kentucky and different levels from the educational system (teachers, administrators, university faculty). Review of the test items by these different groups helped to ensure that the items reflected the Core Content, and that they were grade-appropriate, accurate and fair. As indicated previously, a concerted effort was made to ensure that the KCCT reflected the Core Content for Assessment, essential to assure content validity.

Chapter 4
Item Analysis

Introduction

Item analyses were conducted for each form of the eighteen grade/subject area assessment instruments of the Kentucky Core Content Tests. In this chapter, we present summary information by grade/subject, as well as information about each of the forms. The information includes mean scores and discrimination indices for each item on the form, and reliability indices for the form. The data are based on the calibration datasets that, as described in the chapter "Scaling, Linking, and Producing Scale Scores," include all students for whom complete data were available from the scoring process at the time calibration took place. The decision about when to calibrate for each grade/subject area was based on the sample available. A summary of these numbers is presented in Table 4-1.

For multiple-choice (MC) items, the mean score is simply the proportion of students who gave a correct response to the item (usually referred to as the difficulty index or p-value), and the discrimination index is the point biserial correlation between the item score and the total raw score on the test. The percent range of omits provides the percent omit rate, from the least-often-skipped multiple-choice item to the most-often-skipped multiple-choice item on each form.

For open-response (OR) items, the mean score is the mean of the students' scores on a scale of zero through four. The discrimination is the correlation between the item score and the total score. The percent range of omits provides the percent omit rate, from the least-often-skipped open-ended item to the most-often-skipped open-ended item on each form. As noted in the footnotes of the table, there was one case in which sub-forms differed due to printing irregularities. Sub-forms were designed to differ only in the pretest items they contained. In grade 8 social studies, two items on Form 5 differed, so there were two subforms that differed by two items for the item analyses.

For all grades and content areas, scale scores ranged from 325 to 800 with a mean of about 500 and a standard deviation of about 50 for the first year of the scale, e.g., for Reading grade 4, the scale began in 1992, and for Reading grade 10, the scale began in 1999. As noted in Chapter 7, under Scoring Tables, Kentucky differentially weights OR and MC items for students' final scores. In this item analysis section, weighted values are not used.

Table 4-1
Numbers of Students Administered Each Test Form

Grade	Subject	Number of Forms	Range of N
4	Reading	6	7955-8177
	Science	6	7953-8173
5	Arts & Humanities	12	3944-4059
	Mathematics	6	7917-8040
	Prac. Living/Voc. Studies	12	3942-4056
	Social Studies	6	7915-8037
7	Reading	6	7860-7969
	Science	6	7852-7967
8	Arts & Humanities	12	3832-3968
	Mathematics	6	7727-7844
	Prac. Living/Voc. Studies	12	3825-3966
	Social Studies	5	7715-7841
	Social Studies Subforms*	2	3856-3868
10	Prac. Living/Voc. Studies	12	3550-3701
	Reading	6	7226-7322
11	Arts & Humanities	12	3256-3382
	Mathematics	6	6579-6711
	Science	6	6568-6702
	Social Studies	6	6551-6683

* Forms 5A and 5B in 8th grade social studies differed on two items, so these subforms were processed in separate item analyses.

Grade 4 Forms

Tables 4-2 and 4-3 contain summaries of the item statistics for the grade 4 reading and science forms, respectively. In these and the similar tables for the other grades, "RS" represents raw score; "SS," scale score; "std. dev.," standard deviation; "p-value," item difficulty index; and "pt. Biserial," the point biserial correlation between a dichotomous item score and the total raw score item out.

The alpha reliability is frequently used to measure internal consistency. This measure is used when both multiple-choice and open-ended items are in a test. The alpha reliability is based on a single test administration and provides reliability estimates that equal the average of all split-half reliability coefficients that would have been obtained on all possible divisions of the test into halves. This measure of reliability is the lower bound of the reliability estimate.

The percent range of omits shows the percentage of all students assigned to the form who left an item blank. To illustrate, for grade 4 reading multiple-choice, the range of omitted items for Form 1 was as low as 0.04% and as high as 0.33% of the 8177 students assigned to that form.

Table 4-2
Grade 4 Reading Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple- Choice (24 items per form)						
p-value range	.32 to .88	.46 to .87	.44 to .85	.50 to .92	.48 to .94	.42 to .90
pt. biserial r range	.20 to .48	.23 to .47	.17 to .49	.12 to .53	.25 to .51	.22 to .49
% range of omits	.04 to .33	.07 to .63	.05 to .39	.09 to .56	.04 to .47	.04 to .54
Open Ended (6 four-point items per form)						
mean range	1.61 to 2.10	1.68 to 1.95	1.58 to 1.90	1.70 to 2.15	1.75 to 2.02	1.70 to 2.11
r with total RS	.51 to .65	.54 to .66	.54 to .63	.50 to .62	.49 to .70	.53 to .62
% range of omits	.15 to .57	.16 to .65	.27 to .51	.20 to .73	.19 to .54	.21 to .68
Mean RS	28.5	27.9	26.2	28.4	29.1	29.1
RS std. dev.	7.7	7.7	7.7	7.3	7.5	7.4
Alpha reliability	.89	.88	.87	.87	.88	.88
Mean SS	545.1	546.3	544.6	543.9	546.9	545.3
SS std. dev.	36.9	35.2	35.9	36.4	39.5	35.7

Table 4-3
Grade 4 Science Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple-Choice (24 items per form)						
p-value range	.20 to .94	.44 to .91	.20 to .92	.43 to .85	.40 to .87	.42 to .98
pt. biserial r range	.20 to .48	.16 to .43	.06 to .44	.23 to .45	.15 to .45	.16 to .44
% range of omits	.01 to .31	.05 to .36	.09 to .35	.02 to .36	.09 to .43	.08 to .38
Open-Ended (6 four-point items per form)						
mean range	1.33 to 1.98	.88 to 2.25	1.06 to 2.16	1.44 to 2.77	1.06 to 2.19	1.12 to 2.14
r with total RS	.37 to .62	.40 to .57	.40 to .55	.48 to .60	.39 to .53	.39 to .48
% range of omits	.16 to .53	.14 to 1.25	.22 to .79	.19 to 1.09	.16 to 1.22	.16 to .69
Mean RS	26.8	24.8	24.5	28.0	24.7	27.0
RS std. dev.	8.3	7.5	7.6	8.4	7.3	7.0
Alpha reliability	.84	.84	.82	.85	.82	.81
Mean SS	541.5	540.1	539.8	542.2	540.8	541.3
SS std. dev.	34.0	32.9	33.9	33.7	33.5	33.9

Grade 5 Forms

Tables 4-4 through 4-7 contain summaries of the item statistics for the grade 5 arts and humanities, mathematics, social studies, and practical living/vocational studies forms, respectively.

Table 4-4
Grade 5 Arts and Humanities Summary Statistics by Form

Form	1A	1B	2A	2B	3A	3B
Multiple-Choice (8 items per form)						
p-value range	.46 to .79	.54 to .80	.45 to .80	.46 to .76	.59 to .76	.42 to .80
pt. biserial r range	.20 to .38	.27 to .44	.24 to .39	.24 to .38	.25 to .34	.22 to .36
% range of omits	.00 to .25	.03 to .30	.03 to .50	.03 to .25	.00 to .28	.08 to .20
Open-Ended (2 four-point items per form)						
mean range	1.94 to 2.14	1.77 to 2.10	1.89 to 2.04	1.81 to 2.16	1.61 to 2.10	1.72 to 2.01
r with total RS	.54 to .55	.48 to .56	.47 to .49	.48 to .49	.42 to .46	.47 to .48
% range of omits	.32 to .81	.23 to .38	.13 to .70	.46 to .68	.18 to .81	.23 to .46
Mean RS	9.5	9.0	9.1	9.0	9.0	9.3
RS std. dev.	3.0	3.3	3.0	3.1	3.0	3.1
Alpha reliability	.67	.70	.66	.66	.65	.64
Mean SS	505.6	503.8	507.0	505.2	504.6	506.1
SS std. dev.	65.6	68.2	67.3	66.3	69.4	71.5

Form	4A	4B	5A	5B	6A	6B
Multiple-Choice						
p-value range	.50 to .82	.33 to .86	.41 to .80	.43 to .84	.56 to .82	.32 to .88
pt. biserial r range	.27 to .40	.21 to .45	.21 to .40	.22 to .36	.23 to .41	.18 to .37
% range of omits	.05 to .18	.03 to .50	.03 to .53	.05 to .18	.05 to .58	.00 to .25
Open-Ended						
mean range	1.79 to 2.07	1.70 to 2.00	1.74 to 1.96	2.21 to 2.33	1.88 to 2.22	.99 to 1.54
r with total RS	.45 to .46	.48 to .49	.47 to .52	.41 to .43	.40 to .45	.42 to .47
% range of omits	.30 to .33	.28 to .55	.30 to .45	.25 to .28	.15 to .33	.68 to .94
Mean RS	9.3	9.0	8.7	9.9	9.4	8.1
RS std. dev.	2.8	3.0	3.0	3.0	2.9	3.2
Alpha reliability	.67	.67	.65	.63	.66	.59
Mean SS	507.1	505.8	504.3	510.4	506.2	503.0
SS std. dev.	66.0	67.5	65.7	75.4	67.8	74.5

Table 4-5
Grade 5 Mathematics Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple-Choice (24 items per form)						
p-value range	.42 to .82	.36 to .84	.21 to .91	.32 to .87	.32 to .83	.40 to .83
pt. biserial r range	.20 to .55	.19 to .59	.16 to .50	.19 to .49	.19 to .51	.18 to .54
% range of omits	.05 to .39	.04 to .28	.05 to .28	.06 to .28	.06 to .25	.08 to .41
Open-Ended (6 four-point items per form						
mean range	.35 to 2.12	1.09 to 1.98	.78 to 2.22	1.38 to 2.55	1.02 to 2.56	1.42 to 1.93
r with total RS	.46 to .68	.43 to .63	.51 to .62	.51 to .71	.53 to .62	.46 to .62
% range of omits	.24 to .92	.14 to .62	.29 to 1.19	.15 to .85	.25 to 1.58	.38 to .78
Mean RS	22.7	23.6	22.6	24.9	25.9	24.8
RS std. dev.	9.8	9.5	8.9	9.9	9.7	9.9
Alpha reliability	.88	.87	.85	.86	.87	.87
Mean SS	550.6	553.9	553.0	554.2	553.6	552.8
SS std. dev.	45.7	45.8	44.8	43.5	45.0	47.3

Table 4-6
Grade 5 Social Studies Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple-Choice (24 items per form)						
p-value range	.40 to .80	.45 to .80	.40 to .89	.31 to .86	.35 to .80	.36 to .82
pt. biserial r range	.21 to .48	.13 to .50	.13 to .42	.13 to .48	.12 to .46	.22 to .41
% range of omits	.04 to .29	.04 to .41	.01 to .28	.08 to .34	.01 to .30	.05 to .39
Open-Ended (6 four-point items per form						
mean range	1.84 to 2.21	1.36 to 2.06	1.75 to 2.12	1.48 to 2.14	1.39 to 2.02	1.51 to 2.01
r with total RS	.48 to .60	.43 to .56	.49 to .63	.45 to .58	.49 to .61	.43 to .60
% range of omits	.34 to 1.48	.25 to 1.34	.11 to 1.16	.14 to 1.21	.20 to 1.62	.11 to .99
Mean RS	26.8	26.0	27.2	26.0	25.9	25.2
RS std. dev.	8.1	8.2	7.3	7.0	7.8	8.0
Alpha reliability	.86	.86	.83	.84	.86	.86
Mean SS	535.9	535.6	536.6	535.0	536.1	534.8
SS std. dev.	38.9	39.5	37.4	38.0	37.9	37.9

Table 4-7
Grade 5 Practical Living/Voc. Studies Summary Statistics by Form

Form	1A	1B	2A	2B	3A	3B
Multiple-Choice (8 items per form)						
p-value range	.55 to .80	.43 to .77	.54 to .80	.48 to .83	.26 to .76	.38 to .80
pt. biserial r range	.16 to .42	.16 to .44	.18 to .41	.27 to .35	.10 to .31	.01 to .35
% range of omits	.00 to .52	.03 to .28	.00 to .28	.03 to .41	.05 to .28	.03 to .28
Open-Ended (2 four-point items per form)						
mean range	1.90 to 2.15	2.07 to 2.09	2.12 to 2.17	2.03 to 2.19	1.57 to 2.04	1.45 to 1.98
r with total RS	.38 to .39	.33 to.40	.46 to .48	.41 to .45	.34 to .43	.35 to .38
% range of omits	.35 to .52	.28 to .40	.18 to .38	.25 to .61	.46 to .61	.25 to .58
Mean RS	9.4	9.4	9.5	9.6	8.3	8.1
RS std. dev.	2.9	3.0	3.0	3.0	2.5	2.7
Alpha reliability	.62	.59	.64	.65	.51	.57
Mean SS	500.1	501.9	501.0	501.0	498.4	498.3
SS std. dev.	70.2	76.3	64.7	69.0	65.6	67.6

Form	4A	4B	5A	5B	6A	6B
Multiple-Choice						
p-value range	.59 to .82	.38 to .81	.46 to .79	.46 to .81	.51 to .80	.56 to .78
pt. biserial r range	.24 to .43	.14 to .32	.19 to .39	.19 to .40	.23 to .36	.20 to .41
% range of omits	.00 to .33	.03 to .20	.03 to .28	.08 to .38	.05 to .28	.00 to .30
Open-Ended						
mean range	1.94 to 2.05	2.18 to 2.45	1.84 to 1.96	2.07 to 2.13	2.13 to 2.19	1.91 to 2.23
r with total RS	.39 to .47	.37 to .41	.38 to .47	.37 to .42	.43 to .47	.39 to .45
% range of omits	.28 to .45	.18 to .55	.38 to .45	.35 to .60	.40 to .85	.35 to .86
Mean RS	9.6	10.0	8.8	9.8	9.6	9.5
RS std. dev.	3.0	2.6	3.1	2.7	3.1	3.0
Alpha reliability	.65	.53	.63	.60	.63	.64
Mean SS	501.8	502.4	500.1	502.0	501.6	501.4
SS std. dev.	70.5	67.5	68.7	65.3	69.7	68.8

Grade 7 Forms

Tables 4-8 and 4-9 contain summaries of the item statistics for the grade 7 reading and science forms, respectively.

Table 4-8
Grade 7 Reading Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple-Choice (24 items per form)						
p-value range	.49 to .87	.18 to .91	.33 to .84	.41 to .92	.34 to .86	.32 to .93
pt. biserial r range	.27 to .48	.09 to .48	-.07 to .44	.20 to .49	.19 to .50	.18 to .46
% range of omits	.05 to .40	.06 to .43	.09 to .34	.03 to .32	.04 to .37	.09 to .37
Open-Ended (6 four-point items per form)						
mean range	2.08 to 2.38	1.73 to 2.23	1.85 to 2.20	1.44 to 2.28	1.89 to 2.19	.49 to 2.27
r with total RS	.58 to .66	.54 to .68	.60 to .66	.58 to .66	.58 to .66	.36 to .66
% range of omits	.31 to .90	.64 to 1.24	.51 to 1.40	.63 to 1.83	.56 to .99	.38 to 2.48
Mean RS	29.9	30.1	28.3	28.5	26.9	27.4
RS std. dev.	8.4	7.9	7.5	7.9	8.7	8.0
Alpha reliability	.89	.88	.86	.88	.89	.87
Mean SS	508.8	508.6	511.4	510.1	509.9	509.8
SS std. dev.	36.4	36.0	34.1	34.9	36.1	35.2

Table 4-9
Grade 7 Science Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple-Choice (24 items per form)						
p-value range	.34 to .85	.39 to .96	.25 to .85	.38 to .86	.36 to .83	.30 to .87
pt. Biserial r range	.09 to .50	.16 to .44	.06 to .40	.21 to .48	.14 to .45	.18 to .47
% range of omits	.13 to .46	.08 to .30	.09 to .24	.03 to .39	.10 to .37	.05 to .43
Open-Ended (6 four-point items per form						
mean range	1.55 to 2.32	1.38 to 2.27	1.31 to 2.26	1.13 to 1.81	1.61 to 2.28	1.22 to 2.10
r with total RS	.49 to .67	.52 to .67	.48 to .61	.40 to .57	.52 to .64	.52 to .59
% range of omits	.62 to 1.69	1.18 to 2.25	.57 to 2.19	.90 to 2.11	.65 to 1.64	.78 to 3.72
Mean RS	24.6	24.3	23.9	22.6	25.5	23.9
RS std. dev.	8.8	8.7	7.9	8.5	8.9	9.1
Alpha reliability	.83	.84	.82	.84	.84	.84
Mean SS	497.6	497.7	498.2	496.4	499.0	498.4
SS std. dev.	34.2	33.8	33.0	36.6	33.4	34.1

Grade 8 Forms

Tables 4-10 through 4-13 contain summaries of the item statistics for the grade 8 arts and humanities, practical living/vocational studies, mathematics, and social studies forms, respectively.

Table 4-10
Grade 8 Arts and Humanities Summary Statistics by Form

Form	1A	1B	2A	2B	3A	3B
Multiple-Choice (8 items per form)						
p-value range	.45 to .81	.49 to .78	.26 to .83	.38 to .84	.43 to .89	.45 to .84
pt. Biserial r range	.25 to .39	.24 to .42	.20 to .41	.09 to .35	.00 to .35	.22 to .40
% range of omits	.05 to .18	.05 to .29	.03 to .75	.00 to .31	.00 to .28	.00 to .31
Open-Ended (2 four-point items per form)						
mean range	1.89 to 2.29	2.11 to 2.18	2.03 to 2.37	1.85 to 2.03	1.78 to 1.95	1.91 to 2.26
r with total RS	.50 to .55	.52 to .54	.51 to .54	.54 to .56	.50 to .53	.50 to .56
% range of omits	.68 to 1.44	1.15 to 1.17	1.21 to 1.90	1.04 to 1.22	1.44 to 2.31	.76 to 1.30
Mean RS	9.1	9.2	9.2	8.8	8.8	9.1
RS std. dev.	3.3	3.3	3.1	3.1	2.9	3.1
Alpha reliability	.69	.69	.65	.65	.58	.69
Mean SS	507.0	512.1	507.9	508.8	507.6	508.9
SS std. dev.	65.5	68.2	64.6	64.2	62.0	63.7
Form	4A	4B	5A	5B	6A	6B
Multiple-Choice						
p-value range	.42 to .79	.49 to .87	.43 to .88	.45 to .86	.40 to .79	.51 to .80
pt. Biserial r range	.23 to .42	.27 to .41	.24 to .36	.17 to .43	.10 to .35	.24 to .39
% range of omits	.08 to .65	.03 to .18	.03 to .28	.00 to .31	.05 to .39	.05 to .68
Open-Ended						
mean range	2.13 to 2.28	1.98 to 2.01	2.02 to 2.07	1.85 to 2.21	1.43 to 2.19	1.94 to 2.22
r with total RS	.52 to .54	.52 to .57	.55 to .61	.52 to .56	.46 to .52	.46 to .48
% range of omits	1.35 to 2.12	1.09 to 2.03	.70 to 1.37	.44 to 1.53	1.27 to 2.73	1.27 to 1.30
Mean RS	9.2	9.3	9.1	8.8	8.2	9.0
RS std. dev.	3.3	3.2	3.3	3.2	3.3	3.2
Alpha reliability	.68	.70	.69	.66	.62	.67
Mean SS	508.6	508.7	511.1	508.4	507.3	509.2
SS std. dev.	65.7	68.1	64.5	64.5	69.3	67.9

Table 4-11
Grade 8 Practical Living/Voc. Studies Summary Statistics by Form

Form	1A	1B	2A	2B	3A	3B
Multiple-Choice (8 items per form)						
p-value range	.52 to .80	.50 to .84	.50 to .77	.56 to .80	.43 to .86	.51 to .84
pt. biserial r range	.25 to .35	.10 to .34	.23 to .46	.21 to .43	.12 to .38	.24 to .49
% range of omits	.00 to .38	.03 to .34	.00 to .77	.03 to .65	.00 to .36	.05 to .29
Open-Ended (2 four-point items per form)						
mean range	1.92 to 2.08	2.11 to 2.14	1.64 to 1.90	1.90 to 2.40	1.64 to 1.88	1.65 to 2.21
r with total RS	.46 to .49	.47 to .51	.51 to .52	.53 to .58	.48 to .55	.52 to .57
% range of omits	1.13 to 1.59	1.04 to 1.12	1.78 to 1.96	.91 to 1.64	.85 to 1.67	.92 to 2.20
Mean RS	9.4	9.3	8.7	9.6	8.6	9.3
RS std. dev.	3.1	3.1	3.3	3.4	2.9	3.3
Alpha reliability	.67	.65	.71	.69	.63	.70
Mean SS	500.5	501.3	499.2	502.2	499.9	502.0
SS std. dev.	62.4	61.5	61.7	62.8	58.1	61.3

Form	4A	4B	5A	5B	6A	6B
Multiple-Choice						
p-value range	.57 to .82	.43 to .87	.52 to .82	.54 to .84	.47 to .91	.54 to .82
pt. biserial r range	.26 to .46	.19 to .48	.16 to .44	.30 to .44	.26 to .42	.19 to .43
% range of omits	.05 to .41	.00 to .50	.03 to .31	.00 to .47	.00 to .47	.03 to .68
Open-Ended						
mean range	1.85 to 2.00	1.61 to 2.01	2.03 to 2.13	2.03 to 2.19	2.10 to 2.30	1.81 to 2.00
r with total RS	.47 to .55	.50 to .53	.46 to .50	.48 to .49	.50 to .50	.47 to .58
% range of omits	1.17 to 1.37	.89 to 1.75	.83 to .91	.73 to .96	1.20 to 1.28	.81 to 2.19
Mean RS	9.2	9.2	9.5	9.9	9.8	9.2
RS std. dev.	3.3	3.3	3.1	3.2	3.1	3.2
Alpha reliability	.72	.68	.67	.71	.69	.70
Mean SS	499.2	501.9	501.9	502.5	503.2	500.5
SS std. dev.	63.5	65.5	62.1	64.6	61.6	62.6

Table 4-12
Grade 8 Mathematics Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple-Choice (24 items per form)						
p-value range	.21 to .77	.21 to .78	.14 to .82	.21 to .79	.29 to .85	.26 to .78
pt. biserial r range	.22 to .50	.12 to .52	.20 to .51	.24 to .51	.12 to .54	-.11 to .52
% range of omits	.06 to .40	.04 to .33	.04 to .21	.06 to .41	.06 to .45	.03 to .44
Open-Ended (6 four-point items per form)						
mean range	.89 to 2.12	1.19 to 2.43	.70 to 1.98	1.15 to 1.94	.70 to 2.13	.97 to 2.34
r with total RS	.62 to .68	.52 to .65	.57 to .66	.53 to .69	.56 to .64	.53 to .69
% range of omits	.66 to 1.58	.86 to 1.86	.31 to 1.59	1.05 to 4.92	.28 to 3.54	.67 to 2.20
Mean RS	22.4	23.0	22.0	21.9	21.4	21.5
RS std. dev.	9.7	9.4	10.0	9.7	8.5	8.7
Alpha reliability	.89	.86	.89	.88	.88	.88
Mean SS	527.3	526.3	526.0	527.6	527.3	529.8
SS std. dev.	43.5	44.3	43.9	43.0	42.7	39.5

Table 4-13
Grade 8 Social Studies Summary Statistics by Form

Form	1	2	3	4	5A	5B	6
Multiple-Choice (24 items per form)							
p-value range	.43 to .89	.39 to .90	.26 to .92	.36 to .90	.33 to .89	.33 to .89	.45 to .87
pt. biserial r range	.21 to .48	.09 to .47	.12 to .47	.16 to .46	.17 to .43	.17 to .43	.21 to .51
% range of omits	.00 to .37	.03 to .37	.03 to .28	.03 to .34	.05 to .47	.05 to .47	.06 to .34
Open-Ended (6 four-point items per form)							
mean range	1.55 to 2.34	1.57 to 2.18	1.71 to 2.18	1.67 to 2.28	1.74 to 2.48	1.74 to 2.48	1.87 to 2.44
r with total RS	.57 to .62	.62 to .66	.55 to .67	.57 to .67	.54 to .68	.54 to .68	.55 to .66
% range of omits	.38 to 1.84	.30 to 1.66	.54 to 1.48	.32 to 2.51	.26 to 1.90	.26 to 1.90	.71 to 1.89
Mean RS	27.1	26.4	26.9	27.4	27.8	27.6	27.5
RS std. dev.	9.2	8.8	8.2	8.8	8.7	8.8	9.3
Alpha reliability	.89	.87	.87	.87	.88	.88	.88
Mean SS	508.2	506.5	506.7	508.7	507.6	507.5	509.4
SS std. dev.	48.5	45.7	45.2	46.0	45.0	46.5	47.7

Grade 10 Forms

Tables 4-14 and 4-15 contain summaries of the item statistics for the grade 10 reading and practical living/vocational studies forms, respectively.

Table 4-14
Grade 10 Reading Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple-Choice (24 items per form)						
p-value range	.38 to .85	.23 to .92	.37 to .88	.22 to .85	.39 to .84	.20 to .88
pt. Biserial r range	.16 to .49	.12 to .45	.22 to .50	.07 to .48	.11 to .51	.20 to .47
% range of omits	.12 to .51	.10 to .51	.12 to .59	.12 to .52	.08 to .55	.14 to .61
Open-Ended (6 four-point items per form)						
mean range	1.45 to 2.18	1.57 to 2.25	1.65 to 2.15	1.71 to 2.43	1.51 to 2.08	.94 to 2.10
r with total RS	.66 to .71	.62 to .68	.68 to .71	.64 to .69	.60 to .69	.54 to .66
% range of omits	1.31 to 2.84	.63 to 2.80	1.27 to 2.53	.79 to 3.74	1.20 to 1.98	.90 to 2.89
Mean RS	24.0	26.8	26.7	27.9	26.7	25.2
RS std. dev.	9.9	8.5	9.6	8.9	8.8	8.5
Alpha reliability	.88	.87	.90	.88	.88	.87
Mean SS	496.7	505.6	502.3	503.0	505.0	501.0
SS std. dev.	63.1	56.6	57.3	57.0	57.1	57.7

Table 4-15
Grade 10 Practical Living/Voc. Studies Summary Statistics by Form

Form	1A	1B	2A	2B	3A	3B
Multiple-Choice (8 items per form)						
p-value range	.40 to .78	.56 to .82	.58 to .77	.51 to .81	.51 to .86	.45 to .83
pt. Biserial r range	.14 to .38	.17 to .39	.19 to .33	.17 to .39	.13 to .40	.21 to .41
% range of omits	.05 to .57	.03 to .22	.00 to .30	.03 to .31	.00 to .50	.06 to .36
Open-Ended (2 four-point items per form)						
mean range	1.96 to 2.10	2.05 to 2.23	1.98 to 2.00	1.88 to 1.91	2.10 to 2.16	1.76 to 1.95
r with total RS	.50 to .53	.49 to .53	.52 to .54	.44 to .52	.44 to .44	.47 to .52
% range of omits	1.35 to 1.76	.67 to 1.95	.90 to 1.97	.76 to 1.32	.96 to 1.74	1.94 to 2.39
Mean RS	8.5	9.4	9.3	9.0	9.7	8.7
RS std. dev.	3.5	3.3	3.1	2.9	2.9	3.1
Alpha reliability	.63	.66	.66	.63	.63	.65
Mean SS	492.5	501.6	500.4	501.8	502.6	499.7
SS std. dev.	69.7	64.7	60.8	61.1	64.1	62.0

Form	4A	4B	5A	5B	6A	6B
Multiple-Choice						
p-value range	.61 to .79	.48 to .77	.51 to .77	.51 to .89	.47 to .83	.55 to .81
pt. Biserial r range	.20 to .41	.22 to .41	.24 to .39	.17 to .42	.16 to .37	.16 to .39
% range of omits	.03 to .45	.03 to .45	.00 to .39	.00 to .33	.00 to .28	.03 to .20
Open-Ended						
mean range	1.81 to 1.99	1.69 to 1.93	2.25 to 2.44	1.96 to 2.31	2.20 to 2.42	2.14 to 2.19
r with total RS	.41 to .51	.50 to .53	.48 to .52	.50 to .51	.48 to .52	.46 to .48
% range of omits	1.20 to 1.86	1.34 to 1.67	.95 to 2.53	1.70 to 1.78	.94 to 2.02	.87 to 1.45
Mean RS	9.3	8.6	9.7	9.6	9.7	9.6
RS std. dev.	3.0	3.1	3.3	3.2	3.0	3.2
Alpha reliability	.65	.66	.66	.67	.62	.64
Mean SS	501.4	501.0	504.3	502.9	508.3	504.4
SS std. dev.	64.3	61.2	68.3	62.5	72.2	67.7

Grade 11 Forms

Tables 4-16 through 4-19 contain summaries of the item statistics for the grade 11 arts and humanities, mathematics, science and social studies forms, respectively.

Table 4-16
Grade 11 Arts and Humanities Summary Statistics by Form

Form	1A	1B	2A	2B	3A	3B
Multiple-Choice (8 items per form)						
p-value range	.22 to .75	.54 to .82	.39 to .90	.43 to .87	.28 to .79	.30 to .73
pt. Biserial r range	-.04 to .39	.19 to .35	.24 to .41	.16 to .39	.05 to .32	.14 to .39
% range of omits	.06 to .33	.09 to .18	.00 to .21	.03 to .43	.00 to .28	.00 to .18
Open-Ended (2 four-point items per form)						
mean range	1.71 to 1.99	1.95 to 1.98	1.95 to 2.17	1.59 to 1.85	1.67 to 2.05	1.65 to 1.84
r with total RS	.54 to .56	.54 to .55	.55 to .58	.59 to .61	.57 to .58	.52 to .58
% range of omits	2.07 to 4.46	2.72 to 3.03	1.25 to 2.44	1.82 to 3.52	1.66 to 3.26	1.71 to 3.42
Mean RS	8.7	9.2	8.5	8.3	7.7	7.1
RS std. dev.	3.3	3.1	3.3	3.2	3.1	3.2
Alpha reliability	.67	.65	.69	.66	.63	.67
Mean SS	509.9	506.4	505.4	506.2	502.8	502.0
SS std. dev.	64.4	63.5	66.8	64.5	62.4	62.2

Form	4A	4B	5A	5B	6A	6B
Multiple-Choice						
p-value range	.22 to .80	.34 to .81	.22 to .81	.46 to .82	.37 to .78	.30 to .73
pt. Biserial r range	.04 to .40	.02 to .40	.09 to .37	.19 to .36	.13 to .33	.14 to .41
% range of omits	.03 to .15	.00 to .43	.06 to .98	.00 to .12	.03 to .25	.03 to .61
Open-Ended						
mean range	1.84 to 1.91	1.65 to 2.18	1.87 to 2.32	2.02 to 2.02	1.84 to 1.85	1.70 to 1.99
r with total RS	.56 to .60	.51 to .55	.58 to .60	.45 to .45	.54 to .55	.54 to .62
% range of omits	2.90 to 4.86	2.92 to 4.02	2.38 to 5.49	2.66 to 2.66	1.75 to 3.53	3.89 to 4.08
Mean RS	8.1	8.3	8.5	7.0	8.0	8.0
RS std. dev.	3.2	3.1	3.4	2.6	3.2	3.4
Alpha reliability	.66	.65	.67	.58	.65	.66
Mean SS	505.9	503.9	502.6	441.1	503.3	505.5
SS std. dev.	62.5	62.5	65.4	54.1	71.4	66.3

Table 4-17
Grade 11 Mathematics Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple-Choice (24 items per form)						
p-value range	.19 to .76	.14 to .76	.15 to .69	.16 to .89	.22 to .86	.27 to .87
pt. Biserial r range	.13 to .46	.12 to .49	.00 to .50	.12 to .42	.20 to .48	.12 to .48
% range of omits	.01 to .42	.09 to .65	.06 to .35	.05 to .36	.05 to .76	.06 to .56
Open-Ended (6 four-point items per form)						
mean range	.61 to 1.89	.49 to 2.04	.93 to 1.64	.47 to 1.96	.64 to 2.10	.81 to 1.57
r with total RS	.59 to .66	.52 to .70	.56 to .71	.57 to .72	.60 to .74	.52 to .73
% range of omits	.83 to 4.71	1.27 to 9.68	1.38 to 4.61	1.12 to 7.42	.67 to 6.42	1.23 to 4.70
Mean RS	18.1	17.1	16.8	18.8	19.0	17.9
RS std. dev.	8.5	8.9	8.4	8.2	9.3	8.6
Alpha reliability	.86	.86	.85	.85	.88	.86
Mean SS	524.7	522.1	523.8	526.2	524.8	523.2
SS std. dev.	52.7	59.4	55.1	52.5	57.9	54.5

Table 4-18
Grade 11 Science Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple-Choice (24 items per form)						
p-value range	.24 to .92	.24 to .81	.22 to .74	.25 to .84	.29 to .91	.22 to .90
pt. biserial r range	.11 to .48	.07 to .41	.10 to .44	.15 to .41	.07 to .49	.05 to .44
% range of omits	.04 to .22	.06 to .29	.05 to .33	.03 to .24	.03 to .29	.05 to .32
Open-Ended (6 four-point items per form)						
mean range	.56 to 2.25	.75 to 1.97	.89 to 1.89	.68 to 1.83	.84 to 1.98	.35 to 2.39
r with total RS	.50 to .60	.49 to .56	.46 to .56	.50 to .63	.51 to .65	.35 to .56
% range of omits	.93 to 5.13	1.43 to 8.79	1.96 to 5.86	2.47 to 5.42	1.05 to 6.20	.96 to 7.73
Mean RS	21.8	21.4	19.0	20.4	22.7	21.7
RS std. dev.	8.0	7.4	7.4	8.0	8.1	6.8
Alpha reliability	.82	.83	.81	.85	.84	.79
Mean SS	537.2	538.6	535.0	537.9	535.0	537.7
SS std. dev.	46.0	45.0	46.6	44.7	43.3	43.6

Table 4-19
Grade 11 Social Studies Summary Statistics by Form

Form	1	2	3	4	5	6
Multiple-Choice (24 items per form)						
p-value range	.41 to .92	.37 to .92	.34 to .91	.42 to .96	.35 to .81	.35 to .88
pt. biserial r range	.26 to .48	.16 to .47	.05 to .50	.20 to .46	.11 to .45	.09 to .44
range of omits	.04 to .28	.08 to .26	.08 to .29	.03 to .27	.05 to .24	.08 to .20
Open-Ended (6 four-point items per form)						
mean range	.96 to 1.97	1.36 to 2.07	1.09 to 1.87	1.34 to 2.42	1.46 to 1.99	.94 to 1.98
r with total RS	.60 to .68	.62 to .68	.56 to .71	.61 to .70	.62 to .70	.47 to .66
range of omits	1.11 to 6.45	1.26 to 4.50	1.98 to 5.68	1.09 to 3.17	2.66 to 5.86	1.05 to 2.70
Mean RS	24.8	25.6	22.6	25.9	24.6	23.2
RS std. dev.	9.1	8.9	8.8	8.7	8.9	8.4
Alpha reliability	.88	.88	.86	.87	.87	.85
Mean SS	536.2	539.6	539.4	538.9	540.6	540.4
SS std. dev.	58.9	57.6	58.2	55.5	57.0	58.1

Summary

By examining the means and standard deviations, the forms of the Kentucky Core Content Tests can be considered comparable to each other. For the content areas that have been tested since 1992, the means are currently above 500 and standard deviations below 50. This is most likely due to the length of time these have been tested in Kentucky and the correspondence of some of the KIRIS content standards to the CATS content standards. For Arts and Humanities and Practical Living/Vocational Studies, the means are about 500 and the standard deviations are greater than 50, most likely due to the lower reliability of these tests.

In addition, the internal consistency alpha coefficient for reading, mathematics, science, and social studies are acceptable for student scores. The alpha coefficients for arts & humanities and practical living/vocational studies are not as high, because they are much shorter tests that cover a broad domain. Due to the decreased score reliability, these scores are used in school accountability and not reported at the student level.

The reliability of the Kentucky Core Content Tests in reading, mathematics, science, and social studies is comparable to what might be expected in commercially available standardized assessments. The reliability of the practical living/vocational studies and arts & humanities tests is somewhat smaller because they are relatively short tests, and perhaps because they tend (by design) to measure more diverse content, as opposed to the other content assessments.

Very few students omitted items from the tests. The percentages of students omitting both multiple-choice and open-ended items provide evidence that the test is a power test of the students' skills, not of their ability to pace themselves through a timed assessment.

Chapter 5
Test Administration

Introduction

Great care is taken to assure standard administration of the Kentucky Core Content Test. Close attention to details is necessary to ensure that a student taking the test in one location has an equal opportunity to succeed as a student at another location. Basic information about the administration of the KCCT is available in the ***District Assessment Coordinator's Implementation Guide (DAC Guide).*** That information will not be replicated here, but the following elements are of special interest in this technical report.

Determining Students for Whom a School Is Accountable

Beginning with 1998–1999 and continuing into the 1999–2000 school year, schools were held accountable for scores from the Kentucky Core Content Test, the Writing Portfolio, and the Alternate Portfolio for all students enrolled on the "accountability date." The accountability date was the first day of the Kentucky Core Content Testing window. Nonacademic data for all students in a school was aggregated and included in accountability calculations as well. Nonacademic data includes attendance and retention rates at all levels, dropout rates at the middle and high school levels, and successful transition to adult life at the high school level.

Collecting Enrollment Information

Students and teachers provided school enrollment information on the scannable Student Response Booklets and, at grades 4, 7, and 12, from the Writing Portfolio Score Forms. The Writing Portfolio Score Form is a perforated sheet attached to the front of the Student Response Booklet; it has the same lithocode tracking number as the Student Response Booklet. This lithocode number was then used as a student's identification number throughout the assessment process.

The enrollment information was verified by Data Recognition Corporation (DRC) in two ways for the Kentucky Core Content Tests. First, the information was cross-checked against the data provided by school staff on the *Principal's Certification of Proper Test Administration* form and the District Assessment Coordinator on the district Transmittal form. These forms documented the number of students enrolled in the school on the accountability date and the number of Student Response Booklets returned. Second, the scanned information was compared to a "Student Accountability Roster" that each school was required to send to Data Recognition Corporation. The roster contained the names of all students enrolled on the accountability date.

For the Writing Portfolio Assessment in grades 4, 7, and 12, Data Recognition Corporation compared the portfolio file and student control file (i.e., information from the Student Response Booklet) to ensure that the school accounted for students using both forms of assessment in the Commonwealth Accountability Testing System. For the Alternate Portfolio Assessment, staff

contracted from the Human Development Institute (HDI) at the University of Kentucky verified participating students.

Exemptions

The student exemptions listed below were authorized by the Kentucky Department of Education (KDE) for the Kentucky Core Content Tests.

- A student with Limited English Proficiency (LEP), i.e., students whose native language was not English and who had been enrolled in an English-speaking school for fewer than two years.
- A foreign exchange student.
- A student medically unable to participate in the assessment program, i.e., a student exempted from testing for medical reasons if a signed doctor's statement was provided.
- A student expelled and coded as XP3 or XE3, i.e., not receiving services as provided for in KRS 158.150(2).

The same exemptions were authorized for the Writing Portfolio and for the Alternate Portfolio.

In addition, the Writing Portfolio required that students be enrolled in a Kentucky public school for at least 100 instructional days prior to the accountability date to be included in accountability. This enrollment could have been in treatment centers, detention centers, or homebound instruction programs. Students from out of state but receiving educational services in Kentucky are not considered to be attending Kentucky public schools. Those students attending schools run by the Department of Defense, "home" schools, or private schools also are not considered to be attending Kentucky public schools.

The term "instructional days" applies to those days the school was in session and students were scheduled for class work, thereby excluding professional development days, holidays, snow days, and weekends. Absences, suspensions, and expulsions (other than XP3 or XE3) were not reasons to adjust an individual student's number of instructional days.

Modifications to Data Files

If conflicts in data were noted during scanning or enrollment verification, District Assessment Coordinators were notified to assist in the resolution of the conflict. If the data discrepancy was not resolved at the district level, the information was forwarded to KDE for resolution. Confirmed data changes were made to the master student data files by Data Recognition Corporation. These changes were "flagged" within the master student data file by Data Recognition Corporation when they affected accountability. Final rosters listing the names and scores of students included in accountability calculations accompanied the schools' accountability reports.

Students who were not exempt and did not attempt the Kentucky Core Content Tests, Writing Portfolio Assessment, or the Alternate Portfolio were assigned the *Novice non-performance* level

in the accountability reports. In these cases, either no Student Response Booklet, Writing Portfolio Score Form, or Alternate Portfolio Score Form was returned to DRC or HDI, or the booklets and forms were returned with blank answer areas. Table 5-1 shows the following classification of students for schools years 1999–2000 and 1998–1999:

- The number of students classified as exempt from the Kentucky Core Content Tests;
- the number of students classified as *Novice non-performance* because DRC did not receive an answer document or the response booklet was blank;
- the number who completed the Kentucky Core Content Tests;
- the number of students participating in the Alternate Portfolio Program.

The exempted "other" category includes students who moved out of state or to private schools and were unavailable for testing. Due to differences in record keeping across years, a small number of additional students were included in the "other" category in order to simplify the tables.

Table 5-1[1]
Students In Each Accountability Eligibility Category

[1] Because the KCCT assessment is spread across grades 4 and 5, grades 7 and 8, and grades 10, 11, and 12, these tables are represented by grade sets, as reported for accountability.

Data for School Years 1999–2000 and 1998–1999

1999-2000 School Year						
Accountability Eligibility Category	**Grades 4/5**		**Grades 7/8**		**Grades 10/11/12**	
	Number	**Percent***	**Number**	**Percent***	**Number**	**Percent***
Eligible: Tested	97,316	98.68	94,669	98.17	121,251	97.59
Eligible: Novice (NP)	83	0.08	336	0.35	1,140	0.92
Eligible: Participating in Alternate Portfolio (AP)	622	0.63	839	0.87	1013	0.82
Eligible: Total*	98,021	99.40	95,844	99.39	123,404	99.33
Exempted: Foreign Exc.	3	0.00	3	0.00	336	0.27
Exempted: Medical	141	0.14	219	0.23	179	0.14
Exempted: LEP	373	0.38	329	0.34	288	0.23
Exempted: Expelled	0	0.00	15	0.02	27	0.02
Exempted: Other	12	0.01	19	0.02	8	0.01
Exempted: Total	592	0.60	585	0.61	838	0.67
Total of Eligible and Exempted Students	98,613	100.00	96,429	100.00	124,242	100.00

*Percentages based on total number of students who are eligible and exempt.

1998-1999 School Year						
Accountability Eligibility Category	**Grades 4/5**		**Grades 7/8**		**Grades 10/11/12**	
	Number	**Percent***	**Number**	**Percent***	**Number**	**Percent***
Eligible: Tested	94,323	98.90	95,787	98.61	123,146	98.64
Eligible: Novice (NP)	21	0.02	22	0.02	128	0.10
Eligible: Participating in Alternate Portfolio (AP)	557	0.58	730	0.75	757	0.61
Eligible: Total*	94,901	99.51	96,539	99.39	124,031	99.35
Exempted: Foreign Exc.	4	0.00	0	0.00	369	0.30
Exempted: Medical	157	0.16	238	0.25	214	0.17
Exempted: LEP	266	0.28	291	0.30%	180	0.14
Exempted: Expelled	0	0.00	16	0.02%	17	0.01
Exempted: Other	41	0.04	51	0.05%	27	0.02
Exempted: Total	468	0.49	596	0.61%	807	0.65
Total of Eligible and Exempted Students	95,369	100.00	97,135	100.00	124,838	100.00

*Percentages based on total number of students who are eligible and exempt

Administration of Kentucky Core Content Tests

Testing for all grades took place April 17 through April 28, 2000. Within this testing window, schools were allowed to set up their own specific testing schedules. The state mandated,

however, that content areas must be tested in the sequence found in the test booklets and that all students in the same grade must be tested at the same time.

Coordinators' and administrators' manuals served as guides to administration. The manuals detailed timing requirements, directions for students, and other considerations for administering the tests and handling materials. *Appropriate Assessment Practices,* a document outlining appropriate behaviors for school personnel during testing, was included in the test administration manual. A copy of this document is available from the Kentucky Department of Education. In addition, *703 KAR 5:080 Administration Code for Kentucky's Educational Assessment Program,* the regulation specifying appropriate assessment practices, was included in the District Assessment Coordinator's Implementation Guide.

Seven content areas were assessed using Kentucky Core Content Tests:

- Reading
- Mathematics
- Science
- Social Studies
- Arts and Humanities[2]
- Practical Living/Vocational Studies[3]
- On-Demand Writing

To give more flexibility for administration, reading, mathematics, science, and social studies tests were each divided into three test sections. Arts and humanities, practical living/vocational studies, and on-demand writing tests were presented in one test section each. Test administration time varied from 45 to 90 minutes per test section based on the subject. Test administration procedures provided students with as much additional time as needed, as long as constructive progress was being made toward completion of the test. Additional time, if needed, was to be scheduled *directly after* the initial test session for the content area.

Students received a test booklet and a separate scannable answer document (Student Response Booklet) in which they recorded their answers to all multiple-choice and open-response questions. There were multiple forms of the test, which were spiraled for even and random distribution in classrooms. Each student used only one form of the test for all content areas tested at his/her grade level.

For accountability purposes, there were six forms of the reading, mathematics, science, social studies, and on-demand writing assessments: Forms 1 – 6. For purposes of field testing, there were two subforms for each of the six forms, 12 forms of the assessment in all (Forms 1A – 6B). Each form included one unique open-response pre-test item and four multiple-choice pre-test

[2] This is administered as a single test.

[3] This is administered as a single test.

items.[4] In the areas of Arts and Humanities and Practical Living/Vocational Studies, the 12 forms of the assessment were used for purposes of both accountability and field testing.

Each school principal completed a *Principal's Certification of Proper Test Administration* to confirm enrollment and testing figures. The principal also certified on this form that testing was conducted in accordance with the instructions provided in the manuals and that all materials were handled in a secure manner.

Shipping and Receiving Procedures

All materials were packaged by school and shipped to District Assessment Coordinators. The materials were sent to district offices in three shipments prior to the beginning of test administration. All district and school personnel were informed that the materials were secure and that all secure materials had to be returned to Data Recognition Corporation at the completion of testing.

District Assessment Coordinators were instructed to package all Student Response Booklets from all schools in their districts within one week after the completion of testing. The contracted shipping company then picked up packaged materials from every district office. Kentucky Core Content Test Booklets were packaged and picked up two weeks after testing was completed.

As boxes were received at Data Recognition Corporation, they were collected in a prescribed location for check-in. Boxes were opened one at a time, and Student Response Booklets were checked in at the school level using the *Principal's Certification of Proper Test Administration* as a reference.

To prepare the Student Response Booklets for handscoring, the cover page (student demographics) and multiple-choice page were separated from the open-response pages. The detached pages were scanned, and the open-response pages left in the booklets were sorted by test form. The booklets were then packaged by form for the handscoring process.

Test booklets and other materials were checked for stray Student Response Booklets or other administration forms mistakenly shipped with them. A barcode on each test booklet was scanned to ensure that all secure test booklets were returned and accounted for. If a test booklet was missing, a procedure was in place to require local district staff to determine that the materials were found and returned, or that reasonable steps were taken to assure that these materials were not left in the district and that no security risk remained. After confirmation that all secure test booklets were returned and that KDE had completed all investigations, the booklets were destroyed. *(As of the date of this printing, DRC has not received permission from KDE to destroy.)*

Administration of Writing Portfolios[5]

[4] There were no *pre-test* items included in the on-demand writing test. These items were field tested through a separate procedure.

[5] Alternate Portfolio administration procedures were similar and are detailed in a separate chapter.

Writing portfolios were administered following the model established in 1997–98. A detailed description of the structure and process of training for development and scoring portfolios can be found in the Accountability Cycle 3 *Technical Report* and in *The Writing Portfolio Development and Scoring Process*.

Teacher Training for Portfolio Development

The training model for portfolio development and scoring followed the pyramidal region/cluster organization established for 1991-92. The state was divided into eight regions, with each region assigned a Regional Coordinator for each accountability grade. Regional Coordinators participated as training Cluster Leaders in their regions and served as members of their respective advisory committees. Cluster Leaders, selected by the District Assessment Coordinators in local districts, were the teachers who trained all participating teachers in "clusters" of 20–25 to develop and score portfolios. Initial training focused on developing and implementing portfolios in the classroom, while later training focused on scoring portfolios.

The first phase of training addressed the development of portfolios, fulfillment of requirements, and state guidelines for the generation of student work. Regional Coordinators were the first to receive this intensive training. Regional Coordinators then provided the same training for Cluster Leaders. Kentucky Department of Education staff, assisted by Regional Coordinators, trained the Cluster Leaders. Cluster Leaders then returned to their local districts to provide the same training to all classroom teachers involved in the development of portfolios. Kentucky Department of Education consultants also conducted training that was telecast by Kentucky Educational Television (KET). Districts were encouraged to record the telecasts, or could order the videocassette directly from KET. *The Kentucky Writing Portfolio Development Teacher's Handbook* and *The Kentucky Writing Portfolio Holistic Scoring Guide* were used during the first phase of training.

Training for Scoring

Teachers were trained on the standards and procedures for scoring writing portfolios. This was accomplished through the same training procedures as for development and implementation. As with the Development and Implementation training, *The Kentucky Writing Portfolio Scoring Teacher's Handbook* and *The Kentucky Writing Portfolio Holistic Scoring Guide* were used during this second phase of training for scoring.

Conclusion

As with any standardized accountability instrument, security is a pressing concern. The information outlined in this chapter and detailed in other cited documents indicates the intense

attention paid to the security of the KCCT. At every stage detailed attention is devoted to preventing unauthorized access to the questions and to preventing retention by individuals of inappropriate records. While occasional inappropriate practices emerge which may result in the reduction of student scores, no major breach of security has occurred to date. The thoroughness and cooperation of the contractors is the essential component in the successful administration of the KCCT.

Chapter 6
Scoring

Introduction

The utmost attention to proper construction and appropriate administrative procedures would be to no avail if the scoring of the examinations were careless, inconsistent, inappropriate to Kentucky's standards, or otherwise ineffective. The following materials are intended to give a picture of the scoring process.

Open Response Questions and On-Demand Writing

The 1999–2000 Kentucky Commonwealth Accountability Testing System open-response questions and on-demand writing responses at grades 4, 5, 7, 8, 10, 11, and 12 required handscoring by Data Recognition Corporation personnel. While the processes of selecting and training scorers, reading and scoring papers, and monitoring scoring remained similar to those carried out in previous years for the KIRIS test, these procedures are described below in detail.

Scoring Personnel

DRC has been scoring Kentucky items since 1995. While the scope of work has varied from year to year, DRC has scored over 4.5 million student responses each year for the past three years. Handscoring began on June 19 and finished on August 14, 2000. During this period, over 4.5 million student responses were scored.

Readers
In order to score all items on time, close to 1,000 readers have been employed each year (see Table 6-1). DRC has a large body of scorers who regularly score Kentucky items and items for other states' assessments. DRC selects readers who are articulate, concerned with the task at hand, and flexible. All readers are hired on the basis of their background in the content areas being assessed.

When selecting readers, top preference is given to readers with previous experience scoring Kentucky items and secondary preference is given to readers with previous experience scoring items for other state assessments. It is important to note that the training and quality control procedures are designed to ensure that all scorers, regardless of experience, are able to score Kentucky responses accurately and consistently. KDE requires that all readers have at least two years of college. This requirement is in-line with the requirements DRC has been given by other state departments of education.

Levels of staffing are listed in Table 6-1. The table also shows the numbers of scorers at each grade level who participated in a previous year's scoring (repeat scorers), as well as the number of training leaders. Table 6-2 shows education level and demographic information for scorers in the 1999–2000 testing year.

Team Leaders
Team leaders are selected from the larger body of scorers who regularly score items for Kentucky and for other states. Team leaders are selected on the basis of their proven scoring accuracy and consistency and on their ability to articulate the proper means of scoring.

Scoring Directors
The scoring director staff that trains and leads the team leaders and scorers has been remarkably stable across the years. There is one scoring director for each grade/subject area. In 2000, most of the scoring directors had over four years of experience in a leadership role for handscoring of the Kentucky assessment.

Content Specialists
The scoring directors are trained, supervised, and assisted by DRC's content specialists. These content specialists all have seven to twelve years of experience in handscoring. Each of DRC's seven Kentucky content specials have been working on the Kentucky assessment since 1995 and have been in a leadership role for the Kentucky assessment for at least five years.

Preparation of Scorer Training Materials

The scoring directors and content specialists for open-response questions in each content area met with the WestEd test developer responsible for that domain. The developer, as a facilitator of Kentucky's Content Advisory Committee (content specific), presented the Kentucky objectives, content guidelines, standards, and background information necessary to understand the objectives being measured. Each group also reviewed the framework of the scoring rubric and the language pertinent to the standard.

After this introduction, the combined group read hundreds of student responses and selected anchor papers—papers that typify each score point in the scoring rubric. The anchor responses were annotated for use in the scoring guide. The scoring guide for each item served as the readers' constant reference.

Once the anchors were established, the scoring directors continued the preparation. They identified sets of training papers, similar to the anchor set, and sets of qualifying responses, which included Spring 2000 examples of student responses that represented a range within each score point. Throughout this process, development staff was available to discuss concerns presented by the scoring directors and answer any questions that they might have. Before training sets were reproduced, the scoring directors met with the developers for a final review of the training materials.

Training the Scorers
WestEd development staff was present to observe the initial sessions when the scoring directors presented the standards to the scorers and to provide additional clarification when needed. The

scoring director then completed the training independently. The scoring director and developer consulted as needed throughout scoring; at the end of the project, development staff and scoring directors met to share information about the process and to offer suggestions and comments for future improvement.

Scorers for each content area were selected for their content expertise and were trained by the scoring directors. The scoring directors first presented background information and an explanation of the scoring rubric. The first set of training papers—the anchor training set—was used to clarify the language of the scoring rubric; each score point was illustrated by several anchor papers. This set became the reference set used throughout scoring. Scorers were instructed to review the language of the rubric regularly as they read actual student responses.

The first training set was similar to the anchor set, but papers were in random rather than sequential order by score point. A second training set was designed to instruct scorers how to identify a range within each score point. After discussing the papers in each set, scorers were asked to assign scores independently to another set of papers. The scores were compared to those assigned by the scoring directors and item developers.

Verification of Quality Results

One of the most critical aspects of the handscoring process is to ensure that results are reliable and provide schools with accurate and consistent information. Therefore, a comprehensive verification process is an integral part of all handscoring sessions. A description of this verification process follows.

Qualifying

Prior to scoring, all scorers were required to demonstrate a pre-determined level of scoring accuracy on sets of student responses whose scores were pre-determined during the preparation of scoring training materials as described above. There were two separate qualifying sets, each composed of 15-20 responses. Readers were required to accurately score at least 80% of the responses in one of the sets in order to stay on the project. Any reader failing to achieve an 80% accuracy rate was released from scoring duties and did not score any Kentucky responses.

Consistency Checks

Scorers were monitored for scoring accuracy and consistency daily. Throughout the scoring period, scoring directors and team leaders monitored the reliability of each scorer by re-reading samples of each scorer's work. Each team leader read approximately one packet[1] per scorer each day.

A second monitoring procedure was a second reading of two percent of the responses. Inter-reader reliability reports based on these "double-reads" were produced daily. Tables 6-3 and 6-4 document the percentage of exact agreement between scores assigned by separate scorers. These statistics indicate a high degree of consistency between scorers.

[1] A packet contains 15 student responses.

Based on these two measures of quality control, scoring directors and team leaders carefully monitored each reader's individual performance and the performance of each scoring group as a whole. This allowed any potential scoring drifts to be quickly identified and rectified through further training (including individual training as needed). Readers failing to maintain scoring consistency were released from their scoring duties and their responses were re-scored as necessary.

Scoring Procedures

Student responses were separated by grade and form and placed into packets of 15 student responses each. All open-ended items and on-demand writing responses were scored holistically using a 4-point scale. Each response was scored by one reader. Two percent of the responses were independently scored a second time in order to measure inter-reader reliability (quality control measures are listed in more detail below).

Readers were divided into rooms by subject and grade. Packets were distributed to rooms by clerical staff.

Non-scoreable responses were forwarded to the Scoring Director to assess whether the non-scoreable code should be assigned.

Readers marked as an "alert" any student responses that indicated administrative irregularities or potential dangers for students. Copies of the "alerted" student work were provided to the Kentucky Department of Education.

These general procedures are described in greater detail below.

1. Readers were seated in pairs at long, rectangular tables. Each reader was assigned a unique ID number.

2. The Scoring Director explained in detail the directions for use of score sheet 1 for the first reading of each packet of student responses and score sheet 2 for the second reading.

3. The student responses were separated by grade and form and placed in packets with pre-printed scannable score sheets for each subject area. A clerk distributed the packets of responses to readers. Readers recorded their pre-assigned identification numbers in the designated position on score sheet 1 for their subject area. The readers read each response in the packet that he/she was trained to score and coded the scores on score sheet 1. When all items for the subject in the packet had been read and scored, the readers placed the packets in their "out" bin. The clerk then took the first reader score sheets to the Technical Coordinator for scanning and re-distributed the packets requiring a second read.

4. The second reader coded his/her identification number in the designated position on score sheet 2. That reader then read each response in the packet that he/she was trained to score

and recorded the scores on score sheet 2. After this second scoring was complete, a clerk gave score sheet 2 to the Technical Coordinator. The packets were then taken to the secure storage area for filing.

5. The Technical Coordinator produced inter-reader reliability reports based on the two percent of responses that were scored by two readers.

Conclusion

Reliability of scores is the most important contribution DRC's Performance Assessment staff provides to the KDE. DRC is proud of the work that has been done on the Kentucky Core Content Test. While site locations have changed over the years in order to accommodate the scoring schedule, personnel and quality of training has been consistent. It is this consistency that ensures accuracy of scores.

Meeting the legislated reporting date, given the large volume of student responses to be scored in a relatively short period of time, has presented DRC with many challenges. Excellent training materials, thorough training of readers, quality control measures, and a dedicated and professional staff have all contributed to ensure these challenges have been met.

Table 6-1
Number of Scorers and Training Leaders at Each Grade

Grade	1998 – 99			1999 – 2000		
	# Repeat Scorers (KIRIS)	Scorers	Training Leaders	# Repeat Scorers	Scorers	Training Leaders

4/5	13	267	27	25	307	29
7/8	96	322	30	23	319	30
10/11/12	102	329	31	99	301	28

Table 6-2
Profile of Scorer Qualifications and Demographics

Background		Number of Scorers					
		1998 – 99			1999 – 2000		
		Grade 4/5	Grade 7/8	Grade 10/11/ 12	Grade 4/5	Grade 7/8	Grade 10/11/ 12
Education	Degrees beyond the Baccalaureate	41	68	73	63	61	60
	Bachelor's Degree	173	209	213	198	213	166
	Associate's Degree	20	25	24	27	25	38
	Two-year college study or equivalent	33	20	19	19	20	37
Demographics	Male	107	129	135	147	154	149
	Female	160	193	194	160	165	152
	Black	10	25	17	15	19	18
	White	243	270	293	282	288	268
	Other	14	27	19	10	12	15

Table 6-3
Inter-Rater Reliability in Scoring of Open-Response Questions
Grades 4/5

	1998 – 1999			**1999 – 2000**		
	Exact Agreement	Adjacent Agreement	Non-Adjacent Agreement	Exact Agreement	Adjacent Agreement	Non-Adjacent Agreement
Reading	81	18	1.0	82	18	0.0

Math	83	16	1.0	85	15	0.0
Science	76	22	2.0	80	19.5	0.5
Social Studies	80	18	2.0	81	18	1.0
Arts and Humanities	79	20	1.0	81	18	1.0
Practical Living	82	18	0.0	83	16	1.0
On-Demand Writing	91	9	0.0	80	20	0.0
Average	81.7	17.2	1.0	81.7	17.7	0.6

Table 6-4
Inter-Rater Reliability in Scoring of Open-Response Questions
Grades 7/8

	1998 – 1999			**1999 – 2000**		
	Exact Agreement	Adjacent Agreement	Non-Adjacent Agreement	Exact Agreement	Adjacent Agreement	Non-Adjacent Agreement
Reading	80	19	1.0	87.5	12.5	0.0
Math	84	15	1.0	84	15.5	0.5
Science	81	18	1.0	81.5	18.5	0.0
Social Studies	92	8	0.0	88.5	11.5	0.0
Arts and Humanities	86	14	0.0	84	16	0.0
Practical Living	79	20	1.0	80	20	0.0
On-Demand Writing	86	14	0.0	80	20	0.0
Average	84	15.4	0.6	83.6	16.2	0.2

Table 6-5
Inter-Rater Reliability in Scoring of Open-Response Questions
Grades 10/11/12

	1998 – 1999			**1999 – 2000**		
	Exact Agreement	Adjacent Agreement	Non-Adjacent Agreement	Exact Agreement	Adjacent Agreement	Non-Adjacent Agreement
Reading	86	13	1.0	84	16	0.0
Math	87	13	0.0	88	12	0.0

Science	68	28	4.0	86.5	13.5	0.0
Social Studies	84	16	0.0	90	9.5	0.5
Arts and Humanities	87	13	0.0	88	12	0.0
Practical Living	80	20	0.0	79	21	0.0
On-Demand Writing	86	14	0.0	90	10	0.0
Average	82.5	16.7	0.7	86.5	13.4	0.1

Chapter 7
Scaling, Linking, and Producing Scale Scores

Introduction

This chapter details the procedures for scaling and linking the 2000 Kentucky Core Content Tests to the previous year's scale and describes how the scoring tables were produced.

Scaling and linking were accomplished using the PARDUX and FLUX computer programs. These software programs were developed at CTB/McGraw-Hill to enable scaling and linking of complex assessment data, such as that produced for the Kentucky Core Content Tests.

PARDUX is designed to produce a single scale by jointly analyzing data resulting from students' responses to both multiple-choice and open-response items. In PARDUX, items are calibrated based on item response theory (IRT), using the three-parameter logistic model (3PL, Lord and Novick, 1968) for multiple-choice items and the two-parameter partial credit model (2PPC, Yen, 1990) for open-response items. PARDUX is also used to link the scales developed by two calibrations through the common-item procedure developed by Stocking and Lord (1983).

Item Response Theory Analyses

A marginal maximum likelihood procedure was used to simultaneously estimate the item parameters under the three-parameter logistic model (3PL, used for multiple-choice items) and the two-parameter partial credit model (2PPC, used for performance assessment items) (Bock & Aitkin, 1981; Thissen, 1982, 1986). These models were implemented using the program PARDUX (Burket, 1995). Under the 3PL model, the probability that a student with trait or scale score θ responds correctly to multiple-choice item j is

$$P_j(\theta) = c_j + (1 - c_j)/[1 + \exp(-1.7a_j(\theta - b_j))]. \qquad [1]$$

In equation [1], a_j is the item discrimination, b_j is the item difficulty, and c_j is the probability of a correct response by a very low-scoring student. The 2PPC model holds that the probability a student with trait or scale score θ, will respond in category k to partial-credit item j is given by

$$P_{jk}(\theta) = \exp(z_{jk}) / \sum_{i=1}^{m_j} \exp(z_{ji}), \qquad [2]$$

where $z_{jk} = (k-1)f_j - \sum_{i=0}^{k-1} g_{ji}$, and $g_{j0} = 0$ for all j.

The summary output is in two different metrics, corresponding to the two item response models (3PL and 2PPC). The location and discrimination parameters for the multiple-choice items are in the traditional 3PL metric, and are labeled *b* and *a*, respectively. In the 2PPC model, *f* (alpha) and *g* (gamma) are analogous to b and a, where alpha is the discrimination parameter and gamma

over alpha (*g/f*) is the location where adjacent trace lines cross on the ability scale. Because of the different metrics used, the 3PL (multiple-choice) parameters b and a are not directly comparable to the 2PPC parameters *f* and *g*, however they can be converted to a common metric. The two metrics are related by b = *g/f* and a = *f* / 1.7 (Burket, 1993). As a result of this procedure, the MC and OR items are placed on the Kentucky scale. Note that for the 2PPC model there are m_j - 1 (where m_j is a score level *j*) independent *g*'s and one *f*, for a total of m_j independent parameters estimated for each item while there is one *a* and one *b* per item in the 3PL model.

Goodness-of-Fit: Goodness-of-fit statistics were computed for each item to examine how closely the item's data conform to the item response models. A procedure described by Yen (1981) was used to measure fit. In this procedure, students are rank ordered on the basis of their $\hat{\theta}$ values and sorted into ten cells with ten percent of the sample in each cell. Each item *j* in each decile *i* has a response from N_{ij} examinees. The fitted IRT models are used to calculate an expected proportion E_{ijk} of examinees who respond to item *j* in category *k*. The observed proportion O_{ijk} is also tabulated for each decile, and
the approximate chi-square statistic

$$Q_{1j} = \sum_{i=1}^{10} \sum_{k=1}^{m_j} \frac{N_{ij}(O_{ijk} - E_{ijk})^2}{E_{ijk}},$$

Q_{1j} should be approximately chi-square distributed with degrees of freedom (*DF*) equal to the number of "independent" cells, 10(m_j-1), minus the number of estimated parameters. For the 3PL model m_j =2, so $DF = 10(2 - 1) - 3 = 7$. For the 2PPC model, $DF = 10(m_j - 1) - m_j = 9m_j - 1$. Since *DF* differs between MC and PA items and between PA items with different score levels m_j, Q_{1j} is transformed, yielding the test statistic

$$Z_j = \frac{Q_{1j} - DF}{\sqrt{2DF}}.$$

This statistic is useful for flagging items that fit relatively poorly. Z_j is sensitive to sample size, and cutoff values for flagging an item based on Z_j have been developed and were used to identify items for the item review. The cut-off value is (*N*/1500 x 4) for a given test, where *N* is the sample size.

Model fit information is obtained from the *Z*-statistic. The *Z*-statistic is a transformation of the chi-square (Q_1) statistic that takes into account differing numbers of score levels as well as sample size:

$$Z_j = \frac{(Q_{1j} - DF)}{\sqrt{2DF}}, \text{ where } j = \text{item } j.$$

The *Z* statistic is an index of the degree to which obtained proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters. These values are computed for ten intervals corresponding to deciles of the theta

distribution (Burket, 1991). The *Z* statistic is used to characterize item fit. The critical value of *Z* is different for each grade because it is dependent on sample size.

As a quality control step, all analyses were carried out by CTB/McGraw-Hill research scientists and duplicated by HumRRO scientists.

Scaling and Equating 2000 Kentucky Core Content Tests to 1999 Scales

In this section, we describe the procedures used to calibrate the 2000 Kentucky Core Content Tests items and transform the scales to a metric equated to that used in the 1999 reports. The original scales that Kentucky had were set with a mean of approximately 500 and a standard deviation of approximately 50 for the first year of Kentucky testing, 1992. For Reading grade 10, Arts and Humanities grades 5, 8, and 11, and for PLVS grades 5, 8, and 10, the scale was set with a mean of approximately 500 and a standard deviation of approximately 50 for 1999, the first year these content areas were either tested or scaled.

Item Calibration Samples for all Grades/Subjects

In order to meet reporting deadlines, the 2000 items were calibrated before item response data were available for all students for whom reports were to be generated. When sufficient[1] data were available on all items for each subject/grade, scaling was carried out using the PARDUX computer program. Table 7-1 displays the numbers of students used for the calibrations by grade and assessment subject in 2000.

Table 7-1
Numbers of Students in 2000 Calibration Datasets

Grade	Mathematics	Reading	Science	Social Studies	Arts & Humanities	Practical Living/ Vocational Studies
4		48535	48507			
5	47780			47769	47698	47654
7		47309	47264			
8	46591			46541	46367	46310
10		43503				43239
11	39710		39671	39575	39390	

[1] Sufficient was defined as the availability of data from all forms of the assessment administered, and generally resulted in the use of about 90 to 95% of the student data.

Calibration and Equating Procedures: Grades/Subjects Equated to 1999 Scales

Scaling and equating of these 18 grade/subject assessments was carried out using the PARDUX computer program. The equating method was based on a common set of items referred to as the anchor items, using the method derived by Stocking and Lord (1983). The decision was made that only multiple-choice items would be used in the anchor set due to timing constraints. Furthermore, the anchor items were all included on one of the six forms in each grade/subject of the assessment. Hence the anchor items were the multiple-choice items included on the linking form.

The steps used were:

- create a file of anchor parameter estimates,
- calibrate the 2000 item response data using PARDUX, and
- calculate the Stocking-Lord transformation constants.

A description of each of these steps follows:

As a first step, the parameter estimates in the untransformed PARDUX metric for the anchor items were selected from the file of all parameter estimates and saved in a separate file. Secondly, these estimates were changed into the transformed Kentucky metric, defined using the 1998 data and the constants listed in Table 7-2. These estimates were then saved as an anchor file. It can be noted that all of arts and humanities, practical living/vocational studies and tenth grade reading have transformation constants of 500 and 50 as these content area scales did not exist before 1999, hence they were given constants of the scale mean and standard deviation.

Table 7-2
1999 Linear Scale Transformation Constants

Grade	Subject	M1	M2
4	Reading	33.36	545.54
	Science	27.75	539.77
5	Arts & Humanities	50	500
	Mathematics	35.33	553.01
	Practical Living/Vocational Studies	50	500
	Social Studies	31.61	537.52
7	Reading	31.34	511.37
	Science	26.40	499.30
8	Arts & Humanities	50	500
	Mathematics	33.91	527.60
	Practical Living/Vocational Studies	50	500
	Social Studies	38.38	506.43
10	Practical Living/Vocational Studies	50	500
	Reading	50	500
11	Arts & Humanities	50	500
	Mathematics	39.85	529.85
	Science	31.11	539.99
	Social Studies	44.41	543.55

In the second step, the 2000 student item response data were calibrated using PARDUX. The resulting parameter estimates, including new estimates for the anchor items, were initially in a theta metric.

The Stocking-Lord procedure was then applied to the two sets of estimates and the multiplicative (M1) and additive (M2) constants were determined that would linearly transform the initial 2000 data metric to the Kentucky transformed metric. These constants were then used to produce reporting results in the final scale metric. The transformation constants are displayed in Table 7-3.

Table 7-3
2000 Linear Scale Transformation Constants

Grade	Subject	M1	M2
4	Reading	31.11	547.14
	Science	25.90	543.42
5	Arts & Humanities	49.46	506.50
	Mathematics	34.95	556.46
	Practical Living/Vocational Studies	47.12	500.61
	Social Studies	31.89	537.80
7	Reading	30.43	510.97
	Science	25.55	500.75
8	Arts & Humanities	47.87	510.53
	Mathematics	33.53	530.77
	Practical Living/Vocational Studies	43.54	501.97
	Social Studies	38.96	510.25
10	Practical Living/Vocational Studies	45.08	503.45
	Reading	50.03	506.43
11	Arts & Humanities	47.42	508.29
	Mathematics	40.47	530.80
	Science	31.81	541.74
	Social Studies	46.60	544.62

Producing the Scoring Tables

For each of the 18 grade/subject combinations, tables that show the corresponding scale score for each weighted raw score on each form were produced, with open-response and multiple-choice items received weights of 2 and 1, respectively. Typically, there were six forms for each grade/subject combination except in the arts and humanities and practical living/vocational studies subjects in which there were twelve forms. For some forms, however, there were differences in one item between the A and B subforms. For those forms, separate tables were computed for the subforms. The procedures for computing the values in the scoring tables are specified in the document, "Computing the Raw Score to Scale Score Conversion Tables for the Kentucky Core Content Tests."

The following steps were required to produce each scoring table.

- The estimates of the parameters for the items on the form were selected from the file of estimates of all items in the grade/subject combination.
- A control file for the FLUX program was constructed specifying the M1 and M2 constants for the grade/subject.
- The FLUX program was started and the control file read in.
- The file of parameter estimates for the form was read in.
- The option to weight the open-response items by two was selected and the total weighted score specified as 24 for A&H and PL/VS and as 72 for all other subjects.
- The weighted scoring table was generated and saved as a text file.

Students' score reports were produced using the values in the scoring tables, which contain the scale score equivalent to each raw score and its estimated standard error.

Weighting of Raw Scores

The Kentucky Department of Education instructed CTB to differentially weight the open-response and multiple-choice items. To do this, CTB differentially weighted these items when scoring tables were produced. The computation of these tables is based on the test characteristic function (TCF, sometimes referred to as the expected score function, ESF) in IRT scaling. This function describes the relationship between the proficiency variable (in scale score units) and the expected raw score. In particular, it is derived such that the expected raw score of an individual can be determined from his/her scale score. Note that the scoring table is designed to yield the inverse, an expected scale score from an observed raw score. This is discussed further below.

The expected score function for a single multiple-choice item is simply the item response function:

$$E(r_j|\theta) = P_{j1}(\theta), \qquad [1]$$

where $E(r_j|\theta)$ represents the expected raw score on item j given the scale score, θ, and $P_{j1}(\theta)$ is the probability of a correct score (score of *1*) given a scale score of θ. Given a student's scale score, the function provides the probability of a correct response, which is the expected score on the item for a student having that scale score.

For a test comprised of n multiple-choice items, the TCF is the sum of the ESFs of the n items:

$$\zeta(\theta) = \sum_{j=1}^{n} E(r_j|\theta) = \sum_{j=1}^{n} P_{j1}(\theta), \qquad [2]$$

and it represents the relationship between the expected number of correct responses on the n items and students' scale scores.

For an open-response item scored on an m_j-point scale (*0* to *m-1*), the ESF is given by

$$E(r_j|\theta) = \sum_{k=1}^{m_j-1} kP_{jk}(\theta), \qquad [3]$$

where $P_{jk}(\theta)$ represents the probability of a student with scale score θ getting a score of k $(k = 1,2,\ldots,m_j - 1)$ on item j. Strictly speaking, [3] could be written as summing from the lowest score, *0*, to the highest score, $m_j - 1$, but note that the term in the expression for *k=0* is zero, so that term is unnecessary.

Note that expression [1] can be considered to be a special case of [3] in which $m_j = 2$ because in this case (a multiple-choice item scored *1* or *0*)

$$E(r_j|\theta) = \sum_{k=1}^{m_j-1} kP_{jk}(\theta) = \sum_{k=1}^{1} kP_{jk}(\theta) = 1 \times P_{j1}(\theta) = P_{j1}(\theta)\ , \qquad [4]$$

which is identical to [1]. Hence the expression in [3] may be used for either multiple-choice or open-response items.

The test characteristic function for a mixture of multiple-choice and open-response items on an *n*-item test thus can be written as

$$\zeta(\theta) = \sum_{j=1}^{n} \sum_{k=1}^{m_j-1} kP_{jk}(\theta). \qquad [5]$$

The expression for the probabilities is somewhat complex. For multiple-choice items it may be written as

$$P_{j1}(\theta) = P(x_j = 1 \mid \theta, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}} \qquad [6]$$

$$(j = 1, 2, \ldots, n).$$

In this model:

$P_{j1}(\theta)$ is the probability of a response of *1* given θ, a_j, b_j, c_j, where the "*1*" (second) subscript on P indicates specifically that we are dealing with the probability of response category *1*,

x_j is the response to the *j*th item of an *n*-item instrument,

θ is the proficiency variable,

a_j is a discrimination parameter of the *j*th item,

b_j is a difficulty or location parameter of the *j*th item,

c_j is the lower asymptote of the ICC of the *j*th item.

Note that for any dichotomously scored (two score points) item such as the multiple-choice items under consideration here, there are two possible outcomes, correct and incorrect. In our notation we denote a correct item score as a "*1*" and an incorrect score as a "*0*". The probability of an incorrect score is simply one minus the probability of a correct score,

$$P_{j0}(\theta) = 1 - P_{j1}(\theta),$$

and we need not represent that probability in our model.

For the open-response items the probabilities of the *k* responses are given by

$$P_{jk}(\theta) = \frac{e^{\sum_{i=1}^{k}(\alpha_j\theta - \gamma_{ji})}}{\sum_{t=0}^{m_j-1} e^{\sum_{i=1}^{t} a_j(\theta - \gamma_{ji})}} \qquad [7]$$

$$\gamma_0 = 0$$

$$(k = 0,1,2,...,m_j - 1)$$

$$(j = 1,2,3,...,n),$$

where there is one α parameter for each item and a γ parameter for each response category except zero, for each item. Because each examinee must receive one of the m_j category scores, the probabilities for a given value of θ sum to *1.0* so the probability for category *0* is simply

$$P_{j0}(\theta) = 1.0 - \sum_{k=1}^{m_{j-1}} P_{jk}(\theta). \qquad [8]$$

Computing the values necessary to create a raw score to scale score table involves inverting the function in [5] so the expected raw score could be computed as a function of the scale score.

Given the complexity of the expressions for the probabilities, the inverse function is extremely complex and requires a numerical method to perform the estimation.

Weighting Sets of Items

For the Kentucky Core Content Tests, the decision has been made that open-response items should receive twice the weight of multiple-choice items in determining student performance. This weighting can be accomplished by inserting a weighting factor into equation [5]. The new equation reflecting the weights is

$$\zeta(\theta) = \sum_{j=1}^{n} \sum_{k=1}^{m_j - 1} w_j k P_{jk}(\theta),$$

$$\begin{pmatrix} w_j = 1 for\ multiple-choice\ items \\ w_j = 2 for\ open-response\ items \end{pmatrix} \qquad [9]$$

where w is the weighting factor (2 for the KCCT).

Weighted Raw Score to Scale Score Tables

Appendix 7-1 exhibits the weighted raw score tables for each grade, content area, and form for the Kentucky Core Content Tests. The reading, mathematics, science, and social studies tables have weighted raw scores up to 72 points. The arts and humanities and practical living/vocational studies tables have weighted raw scores up to 24 points.

The lowest scale score (LOSS) for each grade, content area, and form is 325 and the highest scale score (HOSS) is 800. In order to maintain a realistic range of scores that did not vary over forms, content areas, and grades, multiple weighted raw scores converted to the same highest obtainable scale score at the tails of the distribution. These scale score values for the ends of the distribution were established based on an examination of the scale score distributions and the standard error of measurement (SEM) functions for all grades and content areas to keep the range consistent. Due to the constant nature of the range of the Kentucky scale (same LOSS and HOSS for all grades, subjects, and content areas), the standard error of measurement at the tails of the distribution can be large.

Also provided is the performance level that corresponds to the scale scores based on both the KIRIS performance levels, and those levels established in the 2000-2001 standard setting process for CATS. NN is the abbreviation for Novice Non-Performing; NM stands for Novice-Middle; and NH stands for Novice-High. The Apprentice-Low performance level is represented by AL; Apprentice-Middle, by AM; and Apprentice-High, by AH. The Proficient level is represented by a P, and the Distinguished level is represented by a D. All performance levels are obtainable on each form.

Tables 7-1 to 7-6 (Appendix 7-1) illustrate the standard errors, SE, for all of the grade/content area tests. All items on a test were used and the open-ended items were weighted twice the multiple-choice items, as they are for each scoring table. The standard errors for the lowest obtainable scale score and the highest obtainable scale score are provided. In addition, the location of the lowest part of the standard error curve is identified and the standard error is provided for that location.

Chapter 8
Standard Setting

It can be argued that the heart and soul of CATS is the four performance levels used to describe the quality of student work. The levels, from lowest to highest, are Novice, Apprentice, Proficient and Distinguished (NAPD). In addition, the first two levels of performance in reading, mathematics, science and social studies have each been subdivided into three levels (Novice non-performance, Novice medium, Novice high, Apprentice low, Apprentice medium and Apprentice high) to better represent student performance. Kentucky law states that all schools shall expect "a high level of achievement of all students." That high level, defined by the Kentucky Board of Education, is the Proficient level.

On June 5, 2001, the Kentucky Board of Education adopted new standards for CATS. While the new standards will not be fully implemented until the first Accountability Cycle of CATS in 2002, an outline of the standard setting process is provided here because a large part of the development took place in 2000. A detailed Standard Setting Technical Report is available from the Kentucky Department of Education.

The approximately 1600 Kentucky teachers who helped develop the standards participated in three different methods to determine the most appropriate performance standards in each of six content areas. This broad, collaborative advisory process involved teachers from every part of the state. The process itself was designed and overseen by the National Technical Advisory Panel on Assessment and Accountability, NTAPAA. The purpose was to produce a set of clear, consistent, agreed-upon recommendations for standards establishing high expectations for student achievement.
As noted, this process used three different standard setting procedures and had the following six steps:

- Development of Draft Performance Descriptors
- Procedure 1 - Contrasting Groups which focused on *students' classroom performance*
- Procedure 2 - Jaeger-Mills which focused on *student work on the KCCT*
- Procedure 3 - CTB Bookmark which focused on *KCCT test items*
- Synthesis step
- Kentucky Board of Education adoption of the teacher recommended standards.

Step 1 was accomplished in two separate meetings, one in December of 1999 and the other in January of 2000. During these meetings, 88 Kentucky teachers convened to develop a set of Draft Performance Descriptors for each content area and grade level assessed by the KCCT. These Draft Performance Descriptors were developed to establish a common beginning for each of the three standard setting methods. In addition, they were developed to provide a common view of Proficient to allow for the synthesis of the three procedures, or more specifically, the synthesis of the three sets of cut-score recommendations resulting from the three procedures. Perhaps more importantly, the Draft Performance Descriptors were developed with the end product in mind, that is, *to assist teachers in aligning instruction with assessment expectations*. Along these lines, the Draft Performance Descriptors, now called Performance Descriptions,

were refined during standard setting (as part of the procedures) to assure congruence between the demands for students as seen in the content/cognitive descriptions and the demands of the actual assessment. These descriptions by grade level and content area can be found on the Kentucky Department of Education's (KDE) website at http://www.kde.state.ky.us/.

Step 2, the Contrasting Groups procedure, took place in April 2000 and involved 951 teachers. Using the same draft descriptors developed in Step 1, participants used the descriptors to evaluate their own students' classroom performance. Student performance on homework assignments, teacher made tests, classroom participation, etc., was evaluated using the draft descriptors. In other words, these teachers used their own professional judgment and the draft descriptors to categorize their students as Novice, Apprentice, Proficient or Distinguished. If the decision to place a student into one of these four categories was too difficult, teachers were allowed to place the student in one of three borderline categories, i.e., Novice/Apprentice, Apprentice/Proficient or Proficient/Distinguished. While the other two procedures involved teachers coming together in a face-to-face meeting, the Contrasting Groups did not. In other words, no "formal" training for participants occurred as did in the other procedures. In addition, while teachers were provided with written directions on how to apply the Draft Descriptors for making their judgments about students, it is possible that eight years of experience with the old KIRIS cut-scores may have contributed to the judgment of teachers.

Step 3, the Jaeger-Mills procedure, took place in October 2000 and involved 312 teachers who came together for a three-day meeting. The main focus for these teachers was actual complete student work in a content area from the Spring 2000 administration of the KCCT. These teachers also used the Draft Descriptors to categorize student work. Teachers categorized 60 sets of complete student work, each set containing responses to 6 open-response questions and 24 multiple-choice questions. Using the Draft Descriptors, teachers systematically placed each set of student work into one of 12 categories, a low, middle and high category for each of the four performance levels (NAPD). Cut-points for the Jaeger-Mills procedure were obtained by calculating the median value for the "high" and "low" categories of adjacent performance levels, and then taking the middle point between these two values. While the Jaeger-Mills procedure worked quite well, more training time would have been desirable. Similarly, more time refining the descriptors would have also been useful. Finally, in some content areas, the assessment may not have allowed students to demonstrate Distinguished performance relative to the draft descriptors. For example, it is difficult for a single item, or even a set of items, to adequately assess the integration of concepts across content areas or to assess the actual use of manipulatives (e.g., equipment used in science or maps for social studies). This latter observation was very important and led to further refinement of the descriptors to assure congruence between the descriptors and the assessment.

Step 4, the CTB Bookmark procedure, took place in December 2000 and involved 290 teachers who came together for a two-day meeting. The main focus for these teachers was KCCT test items from the Spring 2000 assessment. Prior to the meeting, for each grade level and content area, a book of items was compiled so that the items were ordered by difficulty based on how well students performed on the items in Spring 2000. Items that were easy for students appeared early in the book, while items that were more difficult for students appeared later in the book. Each of the booklets contained both open-response and multiple-choice items. Once again,

teachers used the Draft Descriptors as a starting point. The task for each teacher was to literally place a "bookmark" within the book to indicate the location where a correct response to a particular question would, in the teacher's judgment, place a student into the next higher performance category. Each teacher placed three bookmarks within a book, one for each cut-point, or put another way, one to denote the transition from Novice to Apprentice, from Apprentice to Proficient and from Proficient to Distinguished. Because in Item Response Theory both test items and test takers are put onto the same numerical scale (i.e., the scale score scale), the three bookmarks placed by each teacher translated into three cut-points. Calculating the median value across the teachers within a grade level and content area provided the cut-points from the CTB Bookmark procedure. Two final points about the CTB bookmark procedure are that teachers were given the opportunity to discuss their recommendations prior to submitting final cut-point values and teachers may have been limited by the fact that only part of the total item pool was available for use in the procedure (only 1/3 of the total assessment item pool could be used to construct the ordered item booklets).

Step 5, the Synthesis step, took place in February 2001 and involved 132 teachers who came together for a three-day meeting. For a teacher to participate in the Step 5 Synthesis, the teacher had to have already participated in one of the previous three procedures. The Synthesis step achieved many important objectives. These objectives are summarized in the following bullets where participants had to:

- Understand what had been accomplished in the first four steps of the standard-setting process.
- Evaluate and discuss the instructional implications of the three standard-setting methods.
- Study the recommended cut-scores within the context of impact data.
- Make a subject/grade-level recommendation for the appropriate cut-scores.
- Discuss recommended cut-scores with other subject areas within the same grade level.
- Discuss recommended cut-scores with other grade levels within subject areas.
- Make a final recommendation with impact data to the Kentucky Board of Education.
- Summarize the instructional implications of the cut-scores, and refine the descriptors to fit the cut-score.

The above standard setting project, which took over 18 months to complete, was unique in that it used three different methods to determine the standards. While in retrospect there were some limitations in each method, all three methods were well implemented and consistent with the design as established by the state's National Technical Advisory Panel for Assessment and Accountability. The data from all three methods were valuable in establishing the final recommendations forwarded to the Kentucky Board of Education. In addition to the specific standard setting steps outlined above, between May 10 and May 28, 2001, more than 3,000 people—2,891 identifying themselves as educators—responded to a Kentucky Department of Education online survey about the standards setting process. Slightly more than 32 percent of the respondents said they were "very comfortable" or "comfortable" with the standards setting

process. Only 16 percent said they were uncomfortable with the process. A total of 3,184 people commented on the process by which the standards were developed and/or reviewed the descriptions and submitted comments for the Kentucky Board of Education. The Board in reviewing the standards considered this input. On June 5, 2001, the Kentucky Board of Education adopted the new teacher recommended standards.

As a final note, one of the more important products, if not the most important product, generated from the standard setting process was a set of Instructional Summaries. In fact, in the Synthesis step, *three* sets of Draft Instructional Summaries were provided to teachers, each set based upon the cut-points derived from one of the three procedures (Contrasting Groups, Jaeger-Mills, and CTB Bookmark). *Using the different sets of Draft Instructional Summaries allowed Step 5 participants to evaluate cut-scores without looking at any other data* (e.g., scale scores, distributions of student scores, etc.). It was not until the final day of the Synthesis step meeting that teachers were allowed to view and discuss actual numbers. The following bullets summarize the most important considerations regarding the Draft Instructional Summaries:

- Gave the Synthesis step a beginning point.
- Were improved upon by teachers during the standard setting process.
- Reflect NAPD performance standards resulting from each of the standards setting methods.
- Content – Using the cut-scores identified by each method, an effort was made to summarize the content of items that located or fell within in each performance level (NAPD).
- Cognitive – Using the cut-scores identified by each method, an effort was made to summarize the cognitive skills associated with each performance level (NAPD).

In conclusion, the new standards are important because they define what Novice, Apprentice, Proficient and Distinguished levels of performance mean. They clarify for teachers, students and parents how the Kentucky Core Content Test evaluates student work, and they explain for students what is expected of them. The final cut scores for each grade and content area are in Tables 8-1 to 8-6. The Kentucky scale ranges from 325 to 800 in all grades and content areas. Each scale was set to have a mean of approximately 500, and standard deviation of approximately 50 in 1999. The mean and standard deviations varied some from grade to grade because of relationships to previous KIRIS scaling.

N/A/P/D CUT-POINTS IN KCCT SCALE SCORE UNITS

READING

Performance Standard Cut-Scores*			
	Elem. School	Mid. School	High School
Nov Non/M	326	326	326
Nov M/H	451	426	411
NOV/APP	514	477	454
App L/M	523	488	482
App M/H	532	500	509
APP/PRO	541	511	537
PRO/DIS	601	561	584

MATHEMATICS

Performance Standard Cut-Scores*			
	Elem. School	Mid. School	High School
Nov Non/M	326	326	326
Nov M/H	472	454	457
NOV/APP	546	518	523
App L/M	556	530	535
App M/H	565	543	546
APP/PRO	575	555	558
PRO/DIS	619	584	592

SCIENCE

Performance Standard Cut-Scores*			
	Elem. School	Mid. School	High School
Nov Non/M	326	326	326
Nov M/H	450	434	458
NOV/APP	512	489	525
App L/M	526	498	537
App M/H	540	508	550
APP/PRO	554	517	562
PRO/DIS	588	540	608

SOCIAL STUDIES

Performance Standard Cut-Scores*			
	Elem. School	Mid. School	High School
Nov Non/M	326	326	326
Nov M/H	458	430	446
NOV/APP	524	482	506
App L/M	531	499	530
App M/H	539	516	553
APP/PRO	546	533	577
PRO/DIS	586	580	621

ARTS & HUMANITIES

Performance Standard Cut-Scores*			
	Elem. School	Mid. School	High School
Nov Non/ Nov	326	326	326
NOV/APP	503	478	491
APP/PRO	575	529	554
PRO/DIS	631	610	598

PRACTICAL LIVING / VOCATIONAL STUDIES

Performance Standard Cut-Scores*			
	Elem. School	Mid. School	High School
Nov Non/ Nov	326	326	326
NOV/APP	460	466	458
APP/PRO	507	520	506
PRO/DIS	588	570	578

*Performance Standard levels refer to: Novice Non-Performance/Medium; Novice Medium/High; Novice/Apprentice; Apprentice Low/Medium; Apprentice Medium/High; Apprentice/Proficient; Proficient/Distinguished. Novice Non-Performance is 325 in all content areas.

Chapter 9
Writing Portfolio Assessment: Scoring and Student Performance

The Place of the Writing Portfolio Assessment in the Commonwealth Accountability Testing System

Since 1993, writing portfolios have occupied a key place in Kentucky's assessment programs, both as a means of assessment that directly taps student work in classrooms, and as a means for supporting educational improvement in classrooms, schools, and districts. Since the contents of the portfolios arise from students' classroom work, the portfolio is the assessment component that most clearly reflects local curriculum and instruction. In concept, students develop portfolios over long periods of time—months and perhaps years. Because students have had the opportunity to revise their portfolio entries with support and feedback from teachers and peers, the assessment portfolio may reasonably be viewed as the students' "best work."

In many respects, writing portfolios make up the portion of the Commonwealth Accountability Testing System that most directly and comprehensively supports educational reform, because of the strong connection to students' classroom experiences and the strong involvement of teachers. For that reason, writing portfolio activities include extensive professional development opportunities, which local schools can employ as a powerful means of supporting teachers' professional development and school improvement. A trainer-of-trainers model is employed to deliver scoring training throughout the state. In addition, regional consultants are available to provide professional development and informal teacher support throughout the year.

Local Scoring

All writing portfolios are scored locally to allow each school to observe all the information included in the portfolios—information that goes well beyond the scoring criteria. Reliance on local scoring requires training and practice as well as alignment between portfolio requirements and local instruction. The portfolio development and scoring process also assumes considerable content knowledge of teachers. Local scoring provides a different model of teacher responsibility and involvement in the accountability system from the model provided by the centrally scored or multiple-choice tests. Although external scoring can provide summary data in the form of scores, local portfolio scoring allows discussion of the best ways of modifying instruction based on assessment data that directly reflect classroom practices. Guidance in analyzing results is provided by KDE consultants to schools in the Portfolio Audit as well as to other schools. Extensive professional development is provided throughout the state to support scoring accuracy and the alignment of instruction with the portfolio assessment criteria.

Local Scoring Procedures

A trainer-of-trainers model is employed to deliver scorer training. KDE personnel train Regional Writing Consultants who, in turn, train writing cluster leaders at regional meetings. The writing cluster leaders provide training for the school personnel who score the portfolios. Writing cluster leaders are trained to provide a three to six hour scorer training session. Although a minimum session of three hours is required, six-hour training sessions are recommended.

All scoring sessions are centered on the materials found in the Kentucky Writing Portfolio Scoring Handbooks (the Handbooks for grades 4, 7, and 12 are available from KDE upon request). These materials include directions for conducting a scoring session, all forms needed for a scoring session, and all of the benchmark, exemplar, and high-end portfolios.

Additionally, all scorer training sessions include viewing a video produced by KDE. A new video is produced each year to address frequent questions and "hot" topics and to provide other pertinent and updated information. KDE also provides copies of "seed" portfolios that are used as training, qualifying, and/or validity portfolios at local scoring sessions.

In addition to providing scorer training, writing cluster leaders are trained on procedures for conducting the scoring sessions. Six options are detailed in the Scoring Handbooks. Pros and cons of each type of session are included. Briefly, these options are:

- **Options 1 and 2 - Double Blind Scoring Models:** Both of these options describe procedures for independently scoring each portfolio twice. Scores that are not in exact agreement are resolved through discussion.
- **Options 3, 4, and 5 – Individual Scoring with Selected Double Blind Scoring Models:** These three options all entail scoring each portfolio once individually, followed by an independent second reading for selected portfolios. Scoring discrepancies on portfolios read twice are resolved through discussion.
- **Option 6 – Individual Scoring Model:** All portfolios are scored individually. This option is not recommended.

While there is no requirement for which option is used, KDE specifically does not recommend Option 6, explaining in the Scoring Handbooks that group sessions are preferable because group sessions allow scorers to receive support and feedback from fellow scorers.

Standardizing the Assessment

The writing portfolio assessment is standardized in the following ways:

- Training: The Kentucky Department of Education and the scoring contractor, Data Recognition Corporation, provide every school district with complete scoring training materials for each accountability grade, including detailed rules for portfolio preparation to ensure that the work in each portfolio has been completed by the student.

- Developing: Portfolio content requirements prescribe the number and types of required entries. Portfolios that are incomplete receive zero points.

- Scoring: All scorers use the scoring guide accompanied by several benchmark and high-end portfolios for determining each score point.

Monitoring the System

Portfolio development and scoring are monitored in several ways. Early in the year, an Administration Code is distributed to all schools describing the limits on a teacher's comments or modifications of a student's portfolio entries. When the school-assigned portfolio scores are submitted to the Kentucky Department of Education, the principal is required to submit a signed assurance statement confirming that appropriate portfolio development practices were observed. The Kentucky Department of Education investigates accusations or complaints of inappropriate practices and applies penalties, if warranted. In addition, each student portfolio includes a statement signed by the student attesting that the student completed all portfolio entries. If plagiarism is discovered at local scoring or during an audit, the entry is "removed," making the portfolio incomplete. Incomplete portfolios receive a score of zero.

Due to the public accountability and high stakes associated with the assessment system, the Writing Portfolio Audit is used to monitor portfolio development and scoring. Audits are formal studies of local scoring accuracy. The Audit has several purposes:

- to monitor accuracy of scoring throughout the system in order to plan statewide training and allocation of resources;

- to correct inaccurate scores assigned locally;

- to verify exceptional score gains.

The Kentucky Department of Education defines the sample of schools to be included in the Audit in a manner that allows the results to be generalized beyond the group of participating schools. All schools that are selected must participate and are required to submit all portfolios for rescoring. Locally assigned portfolio scores and the resulting Writing Portfolio Indexes are changed as a result of the Audit.

After eight years of portfolio assessment, two main issues are still associated with the use of portfolios in Kentucky's assessment system: the level of scoring accuracy achieved by Kentucky teachers and the impact of portfolios on instructional practice. The following sections present the rationale and design of the writing portfolio assessment, information about the scoring reliability and instructional impact of writing portfolios during the 1999–2000 school year,[1] and a discussion of related issues.

[1] Information about activities occurring prior to this may be found in the technical report for the corresponding years.

Rationale and Design of the Writing Portfolio Assessment

The Kentucky Writing Portfolio assesses student writing directly (at grades 4, 7, and 12) by examining a collection of a student's written products. The structure of the writing portfolio and the holistic scoring guide encourage teachers to provide instruction focused on teaching students to communicate effectively and to provide grammar, punctuation, and spelling instruction through these authentic writing experiences.

A committee of Kentucky English/Language Arts educators originally designed the portfolio. This committee discussed the traditional writing experience of Kentucky students and discovered that most instruction had focused on isolated grammar and very confined writing experiences (i.e., reports, essays, research papers). Using the writing Academic Expectation (which states that all students should write for multiple purposes in multiple forms for a variety of audiences) as their guide, the committee structured the contents of the portfolio to include broad categories of writing that excluded reports, academic essays, and research papers. However, those categories will continue to be included in instruction. Instead, the committee created a structure that required the following other types of writing.

- Reflective Writing
- Personal Expressive Writing
- Literary Writing
- Transactive Writing

In addition to this purposeful design of the portfolio contents, the criteria for assessment were selected and scoring tools were designed with these instructional focus changes clearly in mind. While the committee believed that mastery and assessment of language mechanics remained critical, they also identified several more critical criteria that had traditionally been less evident in writing instruction and assessment in Kentucky (e.g., focus on real-world purposes and audiences, idea development, and organizational skills).

Finally, the committee selected the following six main criteria for assessing the quality of student writing.

- Purpose/Audience Awareness
- Idea Development/Support
- Organization
- Sentence Structure and Variety
- Language (Word Choice and Usage)
- Correctness (Spelling, Punctuation, and Capitalization)

These criteria were analyzed holistically to produce a single final judgment for each complete portfolio. The committee believed that these portfolio content requirements and assessment

criteria would provide teachers with guidelines for more balanced writing instruction, consistent with the national movement toward more process-centered instruction.

Professional Development

The Kentucky Writing Program (KWP) supports a wide variety of professional development experiences including portfolio scoring training and workshops and consultation focused primarily on classroom strategies for developing student writing skills. Since the introduction of the writing portfolio assessment, a tiered training system has supported classroom teachers. This system relies on a design committee to train local trainers who then deliver portfolio development and scoring strategies to the other teachers in their school. Each year, these local trainers receive two series of professional development training, one focused on the generation of portfolio entries and the other on scoring to state standards. These sessions are augmented by printed materials and video training made available through statewide educational television, and local workshops offered by the writing portfolio regional consultants.

In addition, regional writing consultants work with local districts and schools upon request to provide individually tailored professional development experiences focused on a variety of topics related to the writing portfolio. Examples of topics include portfolio analysis, technical writing, personal expressive writing, reflective writing, writing across the curriculum, development of writing workshop classrooms, and designing appropriate assignments focused on real-world purposes and audiences.

Writing Portfolio Scoring Audit History

The 2000 Writing Portfolio Scoring Audit was the sixth to be carried out. The first Writing Portfolio Scoring Audit was held in 1993 (for details see the Cycle I Technical Report). By legislative directive, the 1993 Writing Audit allowed schools the choice of keeping their original scores or accepting revised scores based on the generally lower audit results. Most schools chose to use the scores that they had assigned to the portfolios to compute their Writing Portfolio Index (WPI).

After the 1993 audit, there were two years of voluntary scoring and analysis sessions. At these sessions, every school with accountable portfolios was offered an opportunity to submit scored portfolios in order to receive analysis regarding their portfolio scoring and development practices.

Beginning in 1996, audits were conducted to monitor statewide scoring patterns and to adjust scores for schools that were scoring portfolios inaccurately. While the 1993 Audit only included schools that were purposefully selected, the 1996 audit included both randomly and purposefully selected portfolios (see “Selection Process” below for more details). Every Audit from 1996 on has used this basic selection model. Those conducting the 1996 Audit observed that local scoring was much more accurate than in 1993. The Audit results were reported to individual schools and were used to adjust the Writing Portfolio Index of all audited schools, as they have been every year since.

After 1996, writing portfolio accountability was moved from grade 7 to grade 8. As a consequence, Writing Portfolio Audits were conducted only at grades 4 and 12 in 1997 and 1998. During these two years, volunteer scoring and analysis sessions were held for grade 7. All schools developing 7th grade portfolios were permitted to submit portfolios to a scoring and analysis session during one of these two years so that every school had an opportunity to receive feedback regarding their scoring practices before being eligible for an Audit.

Several important changes in the writing portfolio assessment occurred beginning with the 1998 – 1999 school year. Beginning in 1999, all three accountable grades (grades 4, 7, and 12) were audited. Also, the calculation of the WPI was altered. Prior to 1999, portfolios that were scored as Novice received zero points on the 140 point WPI scale. Beginning in 1999, portfolios scored as Novice received 13 points on the WPI scale. This change allowed for differentiating between Novice scores and Blank/Incomplete scores by treating Blanks and Incompletes as "Novice-Low" (i.e., zero points on the 140 point index) versus treating the Novice score as "Novice-Medium" (i.e., 13 points on the 140 point scale).

Additionally, in response to legislative directives, the required number of entries in each portfolio was reduced beginning with the 1998 – 1999 school year. Prior to this year, all portfolios at all grades were required to contain a total of six entries of student writing. Beginning with the 1998 – 1999 school year, grade 4 portfolios held four pieces, and grade 7 and 12 portfolios held five pieces. There have been no notable changes since 1999.

Despite the changes outlined above, the audits have, overall, validated the accuracy of local scoring every year from 1996 to the present. The key to this important measurement of stability lies in some critical elements of consistency in the writing portfolio program.

To begin with, there has been great emphasis on maintaining, as much as possible, consistent scorer training materials. The scoring guide, for example, has seen minor changes in the form of dropped Instructional Analysis elements and in enhanced scoring criteria descriptors, but the most important piece, the cell descriptors, has remained unchanged. Likewise, most of the training portfolios (benchmarks, exemplars, and high-ends) have been in place for years, despite the fluctuations in accountable grades and in the required number of entries.

Finally, there has been a consistent emphasis on professional development that has been delivered through the tiered training system from KDE and DRC through the regional writing consultants to the writing cluster leaders and to educators in every school in the state. Together, these key pieces of consistency have translated into a highly stable assessment as measured by the results of the Writing Portfolio Audit.

Writing Portfolio Audit: Rationale, Design, and Procedures

Objectives

Due to the public accountability and high stakes associated with the assessment system, the Writing Portfolio Audit is used to monitor portfolio development and scoring. Audits are formal studies of local scoring accuracy. The audit has several purposes:

- provide a broad picture of statewide scoring accuracy,
- provide data to inform necessary training,
- encourage schools to attend to the accuracy of their scoring,
- ensure that discrepant scores are adjusted,
- establish an environment where auditing is a regular occurrence within the system.

To accomplish all of these objectives, a combination of purposeful and random selection of schools was employed. This type of selection process allows KDE to address the concerns, recommendations, and needs of a variety of audiences (past audit participants, Kentucky scoring teachers, district- and school-level administrators, and external review experts) while retaining equity for all schools.

Selection Process

Since 1996, KDE has identified schools to be audited using a two-stage selection process. This two-stage selection process has changed little, with only minor adjustments being made to the purposeful selection criteria and to the number of schools that were randomly selected. Since 1999, all schools that developed accountability portfolios and submitted scores have been eligible for selection.

Purposeful Selection Model

- A selection index was generated for all schools using current assessment data.
- Schools were rank-ordered based on this selection index.
- Those schools with the highest scores on the selection index were selected for the purposeful sample.
- Schools could not be included in the purposeful sample two years in a row. However, a school may have been part of the purposeful sample one year and still be included in the random sample the following year.

Random Selection Model

After the purposeful selection process was completed, the remaining schools were selected at random. This process ensures that any school may be selected for auditing regardless of the results of the ranking process.

For the 2000 Audit, 106 schools were selected. There were 56 selected at random (referred to as the Random Schools), providing a sample of schools from which to infer statewide scoring accuracy rates. The remaining 50 (referred to as the Purposeful Schools) were chosen using a

formula that identified schools with writing portfolio scores that were very high or very low, relative to test scores in other content areas.

Submitting Portfolios

After purposeful and random selections of schools have been completed, schools were notified of their inclusion in the Audit (May 2000). Selected schools shipped to DRC all portfolios for which original scores were assigned. DRC was responsible for the proper care of all portfolios after they were received and until they were shipped back to schools. To ensure against damage or loss of portfolios during shipment, selected schools were required to photocopy all portfolios before shipping the originals. Schools were reimbursed for any costs incurred for photocopying or shipping. **No photocopies of portfolios were accepted.** Thus auditors were able to score the same material that was originally scored by the teachers.

A score point of zero was assigned to any portfolios which were not submitted but for which scores had originally been reported.

Location and Scoring Team

The 2000 Writing Portfolio Audit occurred in Minnetonka, Minnesota, using DRC's professional writing scorers. These scorers were selected from the larger scoring team that regularly scores on-demand writing responses from the Kentucky assessment, as well as writing assessments for other states. Most members of this scoring team have been professional writing scorers for three to nine years. The team included college graduates, former classroom teachers, educational administrators, writers, editors, retired business people, and other professionals. Scorers were selected based on their demonstrated level of experience and accuracy. In addition, all scorers were required to undergo a process of qualification in order to score audit portfolios. The qualifying procedure, discussed below, is the same as that employed when Kentucky teachers participate in large-scale scoring activities.

Training Procedures

The Writing Portfolio Consultants from KDE and DRC trained all scoring directors, team leaders, and readers using the same procedures and materials used to train all scoring teachers in Kentucky during the school year. The training materials used included the same Holistic Scoring Guide and the "Writing Portfolio Scoring: Teacher's Handbook" used by educators scoring portfolios in-state. In addition, the DRC Writing Consultant trained team leaders and scoring directors in operational and documentation procedures. KDE personnel monitored the auditing session to ensure that the quality of both the scoring accuracy and operational procedures was maintained throughout the process.

Scoring Procedures

Portfolios were packeted within grade levels. Scannable score sheets with pre printed student lithocodes were created for each packet. Scorers provided a score for each portfolio.

As the packets of portfolios were scored, the readers' score sheets were scanned to compare the Audit score to the original score assigned by the school. If these scores agreed, the original score stood as the score of record. If these scores did not agree, the portfolio was scored a second time. After the second Audit score was assigned, all three scores (original, first Audit, and second Audit) were compared. Any two of the three scores which agreed stood as the score of record. If all three scores differed, the portfolio was scored by a scorer of record (KDE/DRC consultants/trainers). Any two of the four scores (original, first Audit, second Audit, and third Audit) which were in agreement stood as the score of record. If there were still no two scores in agreement, the portfolio would be scored by a final scorer of record, a KDE consultant, after which, all previous scores given to the portfolio would be reviewed, and retraining of any readers would immediately occur if deemed necessary. **It is important to note that scorers were not aware of any previously assigned scores (original scores assigned by the schools or scores assigned by other auditors).**

Two points regarding this "reading-to-resolution" model are worth specific mention. First, the process gave equal weight to the original score provided by the schools as it gave to audit scores. In essence, the original scores served as a "first read" for the reading-to-resolution process. Second, portfolios were only read by a second audit reader if the original score and the first audit score were not in exact agreement. In other words, a portfolio only received a second audit read if the portfolio proved to provide scoring difficulties for either the original scorer or the first audit scorer.

Verification of Quality Results

One of the most critical aspects of the auditing process is to ensure that results are reliable and provide schools with accurate and consistent information. Therefore, a comprehensive verification process is an integral part of the audit. A description of this verification process follows (see Table 9-1 for results).

Qualifying

Prior to scoring, all scorers were required to demonstrate a pre-determined level of scoring accuracy on sets of portfolios whose scores had been pre-determined by the Kentucky Writing Advisory Committee and/or the Scoring Accuracy Assurance Team (the standards-bearing sub-committee of the Writing Advisory Committee). Those scorers who successfully qualified began scoring. Those scorers who did not successfully qualify were released from scoring obligations.

Consistency Check

Team leaders (DRC's lead scorers) read behind 20% of the portfolios in every set read by scorers. If scoring discrepancies were noted, discussion and resolution occurred immediately. Scores assigned by both the scorer and the team leader were documented to check against

original scores and to determine the internal level of agreement between scorers (consistency). KDE consultants and DRC scoring directors read behind team leaders conducting the same kind of consistency check. The results of the 20% Consistency Check verify the overall consistency in scoring demonstrated over the span of the audit.

Audit Review

In addition to the consistent monitoring by team leaders, a group of experienced Kentucky scorers, who met the qualifying standards required of DRC readers, were present during the beginning of the audit to conduct the Audit Review Scoring. This team was selected from a large group of teachers who have participated in a variety of statewide scoring activities and have demonstrated consistently high levels of scoring accuracy. Audit Review scorers were trained and qualified to score in the same manner as audit scorers. Portfolios scored in the Audit Review included a random sample of 20% of all audited portfolios. The Audit Review confirmed the quality of the audit scoring.

Accuracy Check: Quality Control Portfolios

Quality Control Portfolios are portfolios whose scores have been pre-determined by the Writing Advisory Committee and/or the Scoring Accuracy Assurance Team, including portfolios that were used in previous audits and have been reconfigured to meet current portfolio configuration requirements. In order to provide continual retraining to Kentucky standards, all readers scored and discussed two Quality Control Portfolios per day. Additionally, the results of the readers' scores were used to verify the consistent application of standards (accuracy). The same procedure was used with the Audit Review Team (Kentucky teachers) to determine the accuracy of the Audit Review.

Reporting Procedures

The following printed information was provided to audited schools:

- A comprehensive document including detailed information about:
  - Training and Scoring Process
  - Audit Results
  - Audit Review Results
  - Overall Quality Results

- Score reports including:
  - Student ID (lithocode number)
  - Original Score
  - Rescore
  - 1999–2000 Writing Portfolio Index

- Cross-tabulation charts including:
  Performance Level Data

Reports and supporting print materials were delivered to Audit schools prior to one-day regional meetings across the state during September 2000. Each school was invited to send three representatives to a regional meeting where they received full information about the quality and results of both the Audit and the Audit Review. Each school had the opportunity to study its results and meet privately (in individual school meetings) with representatives of KDE and DRC to discuss results, patterns, and productive uses of Audit information. Regional writing consultants were present to help schools plan follow-up activities. In addition, Audit Reviewers (Kentucky teachers) were available to discuss their impressions of the quality of the Audit process. It was necessary to hold these regional meetings over a number of days in order to provide to the needs and concerns of each school by making it possible for the greatest number of personnel to attend the meetings. All Audit data provided to schools in the regional meetings was embargoed until all regional meetings had taken place.

Changes to the Accountability Index

When all Audit and Audit Review procedures were completed and the results of the Audit were verified, the scores assigned during the Audit were used to calculate the Writing Portfolio Index (WPI) and the Writing Cognitive Index for all audited schools. All adjustments in scores were reflected in this index. For example:

> School A may have submitted 150 portfolios. The Audit and the Audit Review may have demonstrated that it is necessary to adjust scores for only 6 portfolios. School B may have submitted 60 portfolios, and the Audit and the Audit Review may have demonstrated that it is necessary to adjust scores for 49 portfolios. While School A has demonstrated a substantially greater level of accuracy than School B, both schools' individual portfolio scores will be adjusted to reflect the accurate scores assigned for each portfolio during the Audit.

It is important to understand the impact that a change in a school's WPI will have on the school's overall Accountability Index. The WPI makes up 11.4% of each school's overall index; however, because indices are calculated on a biennium model (two years of data are merged), a single year's WPI makes up only 5.7% of each school's biennium index. Therefore, a ten-point reduction in a school's WPI would result in only slightly more than a half-point decrease in a school's 140-point-scaled biennium index.

The results of the 2000 Audit are summarized in Table 9-2. Two measures of scoring accuracy are presented: the exact agreement between the portfolio scores assigned by the school and those assigned by the Audit, and the magnitude of the difference in the Writing Portfolio Index, a scale of 0 to 140 points. The Writing Portfolio Index is determined by assigning a score of 0 for all portfolios scored "Blank" and "Incomplete." Values of 13, 60, 100, and 140 points are assigned to portfolios rated as Novice, Apprentice, Proficient, or Distinguished, respectively. The Writing Portfolio Index is then computed as the arithmetic mean of all portfolio scores. Note that, since

the values of Novice and Apprentice portfolios were raised beginning with Cycle 4, the data presented below cannot be directly compared with the results of previous Audits.

In the Purposeful group, 73.76% of the portfolio scores were confirmed by one or more Audit readers. For the Random group, 73.32% of the locally assigned scores were confirmed. The Audit adjusted the Writing Portfolio Index for the Purposeful group downward by 9.50 points, while the change for the Random group was a reduction of 8.68 points on the 140-point Writing Portfolio Index.

Together, these two pieces of information, the Writing Portfolio Index changes and the percents of exact agreement, validate the use of local scoring of writing portfolios by verifying the accuracy of local scoring.

It is important to note that scoring accuracy was not uniform across schools. Seventy-eight of the 106 schools audited in 2000 demonstrated a 70% or greater rate of scoring agreement with audited scores. Table 9-3 shows the variety of scoring accuracy found among audited schools over the last three years.

These data are further detailed on cross-tabulations available from KDE upon request.

Selected Reports on Writing Portfolios Available from the Kentucky Department of Education

- 1992–1993 Writing Portfolio Rescoring Report
- 1993–1994 Writing Portfolio Scoring Analysis Report
- 1994–1995 Writing Portfolio Scoring Analysis Report
- 1995–1996 Writing Portfolio Audit Final Report
- The 1996-97 Writing Portfolio Audit: Rationale and Procedures
- The 1997-98 Writing Portfolio Audit: Rationale and Procedures

Portfolio Issues Summary

Several trends emerge from the data presented on Tables 9-2 and 9-3. When viewed across the years, the data present an interesting pattern in scoring, which might be summarized as, “instruction makes perfect.” Teachers in the accountability grades score many writing portfolios each year and their agreement with external scorers is generally greater than 70%. Every year,

they are also faced with the challenge of adapting their instruction to the portfolio-scoring criteria.

Within these acceptable levels of overall agreement, there are two interesting patterns. First, it is notable that for the last two years, the agreement rate at grade 12 has been lower than the agreement rates at grades 4 and 7. However, because there have been fewer grade 12 schools included in the Audit, it is difficult to know if these differences are truly a reflection of local scoring trends across the state.

A second pattern is that the overall agreement rates for the randomly selected schools have been more stable across the years than the rates for the purposefully selected schools. In 2000, the agreement rates for the two groups were virtually the same. This trend may indicate a need to re-visit the purposeful selection criteria.

Finally, it is worth noting that locally assigned scores are consistently most accurate at performance levels that occur most frequently. As one observer noted, "After you have scored 300 Novice portfolios, you have an idea of where Novice begins and ends. And after reading 300 Apprentice portfolios, you have an idea of what they look like. But probably no one [local scoring team] in the state has yet scored 300 Proficient portfolios. It's really not surprising that accuracy is a little lower there." The implications for ongoing scoring training are clear. To increase the accuracy of identifying true Proficient and Distinguished portfolios, portfolio scorers need many opportunities to read and score these high performance level portfolios. The challenge is to find and distribute more examples of student work at these performance levels.

Table 9-1
2000 Audit Quality Control Results

Audit Team Leader to Audit Reader % Agreement	Audit Reader to Validity Portfolios % Agreement	Review Team to Final Audit % Agreement
83.91%	91.18%	76.66%

Table 9-2
Summary of Audit Results

		PURPOSEFUL SCHOOLS			RANDOM SCHOOL		
YEAR	GRADE	ORIGINAL AUDIT SCORE AGREEMENT	ORIGINAL WPI	AUDITED WPI (DIFFERENCE)	ORIGINAL AUDIT SCORE AGREEMENT	ORIGINAL WPI	AUDITED WPI (DIFFERENCE)
2000	GRADE 4	76.92%	68.65	61.70 (-6.95)	80.18%	57.52	54.80 (-2.72)
	GRADE 7	71.53%	40.71	28.96 (-11.75)	76.41%	42.93	34.76 (-8.17)
	GRADE 12	67.76%	66.90	54.31 (-12.59)	60.12%	65.35	48.44 (-16.91)
	SUMMARY	73.76%	57.10	47.60 (-9.50)	73.32%	53.79	45.11 (-8.68)
1999	GRADE 4	77.80%	70.89	71.59 (+.07)	77.84%	58.35	57.50 (-.85)
	GRADE 7	70.65%	52.54	40.65 (-11.89)	78.13%	35.13	28.10 (-7.03)
	GRADE 12	48.14%[1]	70.72	47.30 (-23.42)	71.97%	60.67	50.06 (-10.61)
	SUMMARY	74.90%	64.76	60.87 (-3.89)	76.37%	49.83	43.90 (-5.93)
1998[2]	SUMMARY Grades 4 & 12	69%	54	40 (- 14)	75%	46	36 (-10)
1997	SUMMARY Grades 4 & 12	77%	46	39 (- 7)	74%	49	40 (-9)
1996	SUMMARY Grades 4, 8, & 12	73%	44	33 (-11)	77%	32	27 (-5)

[1] In 1999, only one grade 12 school was purposefully selected for the audit.

[2] Prior to 1999, the WPI was calculated using 0 points for Novice scores. Beginning in 1999, WPIs were calculated using 13 points for Novice scores. Therefore, the WPI data prior to 1999 cannot be compared with later WPI data.

Table 9-3
Levels of Exact Agreement

PERCENTAGE OF EXACT AGREEMENT	NUMBER OF SCHOOLS				
	1996	1997	1998	1999	2000
90% or greater	10	16	8	4	6
80–89%	33	30	31	41	39
70–79%	25	34	31	33	33
60–69%	18	9	13	16	16
50–59%	8	4	10	7	5
Less than 49%	4	7	7	5	7

Chapter 10
Alternate Portfolio Assessment

Rationale and Participation Guidelines

The Kentucky Alternate Portfolio (KAP) is the assessment vehicle for students with disabilities who cannot participate in the regular Commonwealth Accountability Testing System, even with the provision of program accommodations or modifications or both. These are students with documented cognitive disabilities typically ranging in the moderate to severe functioning levels.

When addressing the issues of student assessment and program accountability, it is important that all students be included in such endeavors. In this effort, the Kentucky Alternate Portfolio was developed and has been implemented since 1991. The Kentucky Alternate Portfolio was developed by a team of Kentucky special education teachers, local school administrators, and Kentucky Department of Education and University of Kentucky staff to reflect educational outcomes that are important for all students, including those with moderate to severe disabilities, and is consistent with the Kentucky Education Reform Act.

Contents

The Kentucky Alternate Portfolio has a set of required elements modeled after the Kentucky Writing Portfolio contents. Each portfolio must include a Table of Contents, a Letter to the Reviewer, a Parent Validation Letter, an Individualized Schedule reflecting the student's mode of communication, and five (5) entries consisting of samples of the student's best work. These entries should reflect instruction over time in content areas. The entry types follow the curriculum advanced in the Kentucky Program of Studies and are grade level specific. Table 10-1 identifies the entries required by grade.

The entries may include work collected over time and should reflect instruction toward the achievement of the Kentucky Learner Goals and Academic Expectations. The Kentucky Alternate Portfolio is specifically based upon a subset of these standards, but may and should show examples of work toward other expectations as necessary. Appendix 10-1 contains a listing of this subset.

Table 10-1
Alternate Portfolio Requirements

A complete portfolio will include the following items:

- A table of contents (written, pictorial, audio/videotape)
- A letter to the reviewer that describes the portfolio contents (written, pictorial, audio/videotaped)
- A letter from the parent validating contents of the portfolio
- An individualized daily schedule with description and documentation of student use
- 8th and 12th grade vocational entries: career exploration (at least 3 relevant activities) and a formal résumé, respectively
- Student mode of communication consistently evidenced throughout
- Five entries from the following areas (official entry cover sheets are required and can be found in Appendix 10-1).

Entry Types	4th Grade	8th Grade	12th Grade
Language Arts	X	X	X
Math	X	X	C
Science	X	C	C
Social Studies	X	C	C
Arts and Humanities	C	C	C
Health and PE	C	C	C
Vocational	N/A	X	X

X = Required
C = Choice
N = Not Applicable

These entry types are directly linked to the Academic Expectations in both Transformations and the Program of Studies. The Program of Studies document is available on KDE's website. This document, coupled with collaboration between teachers in the building, should yield a wealth of ideas for entries that also support the participation of non-disabled peers. TASKS, an additional support document, can also be found on the web at http://www.ihdi.uky.edu/products . TASKS highlights the multiple ways that students with diverse learning needs can evidence performance of the academic expectations while participating in regular education activities directly related to the Program of Studies. Both resources should be helpful in developing and implementing instructional activities that can be used in portfolio entries.

Scoring

Kentucky Alternate Portfolios are scored for accountability at grades 4, 8, and 12. This may have a slight variation since the student himself/herself is included in the accountability cycle as a result of age. A student in alternate assessment is considered to be a fourth-grader if he/she is 9 or 10 on October 1 (but no older than 11), an eighth-grader if he/she is 13 or 14 on October 1 (but no older than 15), and a twelfth-grader when he/she is 18 on October 1 or in the last anticipated year of school.

Kentucky Alternate Portfolios are scored holistically according to a rubric that is described in Table 10-2. The rubric reflects best practice instruction in special education with criteria regarding student progress, self planning/monitoring/evaluation, work toward standards set for all students, multiple settings for instruction, support, social relationships, and student choice.

Table 10-2
Alternate Portfolio Scoring Rubric

	Novice	Apprentice	Proficient	Distinguished
Performance	Student participates passively in portfolio products. No clear evidence of performance of specifically targeted IEP goals/objectives. Little or no linkage to Academic Expectations.	Student performs specifically targeted IEP goals/objectives that are meaningful in current and future environments. Planning, monitoring and evaluating are limited or inconsistent. Some evidence of Academic Expectations.	Student work indicates progress on specifically targeted IEP goals/objectives that are meaningful in current and future environments. Student consistently plans, monitors, and evaluates his/her own performance. Academic Expectations clearly evidenced in most entries.	Student work indicates progress on specifically targeted IEP goals/objectives that are meaningful in current and future environments. Planning, monitoring, and evaluating progress is clearly evident. Evaluation is used to extend performance. Extensive evidence of Academic Expectations in all entries.
Settings	Student participates in limited number of settings.	Student performs targeted IEP goals/ objectives in a variety of integrated settings.	Student performs targeted IEP goals/ objectives in a wide variety of integrated settings within and across most entries.	Student performance occurs in an extensive variety of integrated settings, within and across all entries.
Support	No clear evidence of peer support or needed adaptations, modifications, and/or assistive technology.	Support is limited to peer tutoring. Limited use of adaptations, modifications, and/or assistive technology.	Support is natural with students learning together. Appropriate use of adaptations, modifications, and/or assistive technology.	Support is natural. Use of adaptations, modifications, and/or assistive technology evidences progress toward independence.
Social Relationships	Student has appropriate but limited social interactions.	Student has frequent, appropriate social interactions with a diverse range of peers.	Student has diverse, sustained, appropriate social interactions that are reciprocal within the context of established social contacts.	Student has sustained social relationships and is clearly a member of a social network of peers who choose to spend time together.
Context	Student makes limited choices in portfolio products. Products are not age-appropriate	Student makes choices that have minimal impact on student learning in a variety of portfolio products. All products are age-appropriate.	Student consistently makes choices that have significant impact on student learning. All products are age-appropriate.	Student makes choices that have significant impact on student learning within and across all entries. All products are age-appropriate.

The initial scoring of Kentucky Alternate Portfolios was done regionally with teams of two special education teachers scoring portfolios from districts other than their own. Each portfolio was scored by two different teams. If the holistic scores of those teams matched, that was the final score assigned to that portfolio. If the scores did not match, the portfolio was scored a third time by another team or a state-certified scorer. The two matching scores were the final, assigned score for that portfolio.

In instances where a teacher did not agree with the final score assigned to his/her student's portfolio, an appeals process was in place. In these cases, the teacher took the portfolio back to his/her district, scored it him/herself, and wrote a rationale outlining the points of disagreement and explaining why he/she thought the portfolio should have been scored differently. The teacher then sent the portfolio to the Kentucky Alternate Portfolio project office where it was assigned to be scored by a state scorer. The score assigned by that state scorer was the final score for the portfolio. State scorers are teachers who have received extra training in scoring alternate portfolios and have demonstrated competency in this area.

Kentucky Alternate Portfolio scores were aggregated into each school's total accountability index, resulting in that school's accountability for all students. The Kentucky Alternate Portfolio score for any specific student was entered into all seven of the assessment areas (reading, math, science, social studies, arts and humanities, practical living and vocational studies, and writing) for that level (elementary, middle, or high school). Table 10-3 illustrates this. This gives schools important information to be used in consolidated planning for instructional improvement, which includes students with moderate and severe disabilities.

Table 10-3
Grade Levels and Content Areas Incorporating KAP

	Reading	Math	Science	Social Studies	Arts and Humanities	Practical Living/ Vocational Studies	Writing
4th Grade	X		X				X
5th Grade		X		X	X	X	
7th Grade	X		X				X
8th Grade		X		X	X	X	
10th Grade	X					X	
11th Grade		X	X	X	X		
12th Grade							X

Monitoring the System

Scoring consistency was based upon the degree of agreement between the two team scores. The alternate portfolios were scored by special education teachers who received training similar to that received for scorers of the writing portfolio. Training consisted of a full day, regional sessions. The morning of the training focused upon clarifications of the rubric and the afternoon was spent in guided practice of scoring benchmark portfolios.

Alternate Portfolio Reliability or Consistency

For the school year 1999-2000, a total of 918 alternate portfolios were submitted for scoring. There were 314 fourth-grade portfolios, 317 eighth-grade portfolios, and 287 twelfth-grade portfolios.

The total numbers of portfolios having initial agreement in scores between the two teams of scorers were 173 for the 4th grade, 174 for 8th, and 156 for 12th. The percentages of agreements in scores were as follows:

- 4th grade—55%
- 8th grade—55%
- 12th grade—54%.

The inter-rater agreement for all grade levels combined is approximately 55% (weighted by number of students at each grade).

The relatively low percentage of inter-rater agreement can be attributed to several factors including the relative "newness" of the portfolio entry types, the total number of scorers, high degree of turnover of scorers (i.e., special education teachers), holistic nature of the portfolio itself, and the number of content areas addressed by the portfolio.

In the 1998-1999 school year, the entry types required in the alternate portfolio was revised. Instead of activity-based entries in which the student's performance was documented in the context of typical instructional activities, entries were required to document the student's performance over time in specific content areas (refer to Tables 9.1 and 9.3). This has required that schools and programs learn to document in a new format and that instruction be even more curriculum or content area based. These are new directions for instruction and assessment for students with moderate and severe disabilities and, again, Kentucky is leading the way by means of its alternate assessment.

Because Kentucky places a high priority on the professional development opportunities (and ensuring possibilities of instructional improvement) available through scoring alternate portfolio assessments, all special education teachers are required to score portfolios if they have students in the current accountability year. This means approximately 600 teachers across the state are trained and score alternate portfolios yearly. A smaller number of trained scorers would

probably raise the percentage of inter-rater agreement but would lower the opportunities for professional development. Some of these teachers may have a student or students in accountability one school year and then may have no one in for several years after that. For teachers in situations like this, the length of time between trainings can be several years, resulting in a pattern of constantly "new" learning. Of those 600 teachers per year, approximately 1/3 are not only new to the portfolio process but are new to teaching. This mirrors the rate of turnover in the field of special education and probably contributes significantly to the low percentage of inter-rater agreement.

The holistic score of the alternate assessment facilitates its aggregation into schools' accountability indexes. However, the nature of holistic scoring procedures can contribute to scoring discrepancies. That, combined with the amount of information required to adequately document student performance in all content areas, all learner goals and academic expectations, and the amount of time potentially covered by the evidence in the portfolio (i.e., up to 3 years), could account for some of the inter-rater disagreement.

The KAP project acknowledges the concern over the low percentage of inter-rater agreement and the lack of progress towards increasing the percentage in previous years. It has been proposed that a series of scoring "experiments" be conducted during the next school year so that method of scoring be established that will increase the inter-rater agreement. These experiments would begin with a thorough review of the literature and then be conducted with a sample of Kentucky special education teachers. The experiments are expected to include teachers from across the state with varying levels of experience and to look at such factors as pair versus single scoring.

Chapter 11
Reliability

Introduction

There are many aspects of the Kentucky Core Content Test (KCCT) reliability that are of concern. Since reward and assistance decisions for schools are based primarily upon the KCCT, it is very important that we have reliable scores. To that end we must have a reliable scoring process. To have a reliable scoring process we must keep the various forms of the Kentucky Core Content Test (KCCT) equivalent within each year by careful form building, and across year by equating the test scores. Probably no issue is of greater concern then that of proper classification for both student, into one of four performance levels (Novice, Apprentice, Proficient, or Distinguished), and school into one of three major performance levels (Meeting Goal, Progressing or Assistance).

Student-Level Reliability

Although accountability decisions apply at the school level, student-level results are reported to parents. It is important, therefore, to examine reliabilities at that level. Table 11-1 presents student-level coefficient alpha for 1997 through 2000 (Interim Accountability Cycle) by grade, subject, and year. For 1997 and 1998 the coefficient alpha is computed by form for common and matrix open-response items combined. Median and range alpha values are computed across all 12 forms of the old Kentucky Instructional Results Information System (KIRIS). For 1999 and 2000 coefficient alpha is also computed by form; however, because Kentucky students were tested using the new Kentucky Core Content Test (KCCT) rather than the old KIRIS there are no within year common items between forms for 1999 and 2000. Median and range alpha values are computed across the 6 forms in four subjects (Reading, Mathematics, Science, and Social Studies) of the KCCT using both multiple-choice and open-response items. All these values are based on data contributed by students who were eligible to complete testing and who were present on the day of testing. Absence is defined as an observation in which all items are blank. When an observation includes at least one response, zeros corresponding to any blank items are entered in the computation of coefficient alpha. The responses of absent students (all blanks) are excluded to avoid overestimation of score reliability.

Table 11-1
Test Reliabilities
Coefficient Alpha by Grade and Subject

Grade	Subject	1997		1998		1999		2000	
		Median[1]	Range	Median[1]	Range	Median[2]	Range	Median[2]	Range
4/5	Reading	.80	.77-.82	.81	.80-.84	.87	.87-.88	.88	.87-.89
	Mathematics	.76	.74-.77	.82	.81-.83	.86	.84-.88	.87	.85-.88
	Science	.72	.71-.76	.71	.68-.73	.82	.78-.83	.83	.81-.85
	Social Studies	.77	.73-.79	.76	.73-.79	.83	.82-.87	.86	.83-.86
7/8	Reading	.85	.82-.86	.85	.84-.86	.88	.87-.89	.88	.86-.89
	Math	.79	.78-.81	.79	.78-.80	.87	.85-.89	.88	.86-.89
	Science	.77	.73-.79	.79	.76-.82	.84	.82-.85	.84	.82-.84
	Social Studies	.85	.82-.86	.85	.83-.86	.87	.87-.89	.87	.87-.89
10/11[3]	Reading	.86	.83-.87	.87	.84-.88	.88	.85-.90	.88	.87-.90
	Mathematics	.82	.79-.83	.86	.85-.87	.85	.84-.88	.86	.85-.88
	Science	.80	.76-.82	.81	.78-.83	.81	.80-.85	.82	.79-.85
	Social Studies	.86	.83-.88	.89	.86-.90	.86	.84-.88	.87	.85-.88

[1] Median coefficient alpha based upon common and matrix open-response items across the 12 forms of the KIRIS.
[2] Median coefficient alpha based upon operational matrix open-response and multiple-choice items across the 6 forms of the KCCT.
[3] Reading in 1997 and 1998 was tested in the 11th grade. In 1999 & 2000 Reading was tested in the 10th grade.

The Kentucky Department of Education (KDE) advises against making student-level decisions based on individual test scores alone. However, both KIRIS and KCCT test reliabilities compare favorably with reliabilities from other tests used in the process of making student-level decisions. Individual KIRIS and KCCT subject area reliabilities are similar to ACT and CTBS subject area reliabilities.

The increase in reliability coefficients from KIRIS (1997 and 1998) to CATS (1999 and 2000) is apparent in all subjects except high school social studies. This increase is a benefit due to increased test length by the introduction of multiple-choice questions into the mix of assessment types, although the main motivation for this addition was to broaden the KCCT content coverage found in the *Core Curriculum for Assessment.*

Note that using coefficient alpha probably underestimates score reliability insofar as item raw scores are the basis for the computation. The fundamental scaling method used with KCCT employs a logistic model. The use of item response theory takes into account differences in item difficulty not reflected in the computational use of raw scores utilized in computing classical test theory reliability estimates found here.

A limit on coefficient alpha is the prospective for irrelevant variability in student-level scores arising from the use of potentially non-equivalent multiple scorers for open-response questions. As indicated previously in Table 6-1, the effect of having different open-response scorers appears minimal even at the student level. The lowest average percentage of exact agreement both in 1999 and 2000 is 81.7% for the KCCT given in the elementary grades. The highest average KCCT percentage of exact agreement is found in 2000 at 86.5% for the high school grades. This good inter-rater scoring agreement, along with the fact that non-adjacent score

points inter-rater scoring of open-response questions was low (between 0.1% and 1.0%, depending on the year and school level), shows that multiple scorers have little negative effect on overall test reliability. In addition, the fact that scoring is monitored through a quality assurance process, suggests that the effect of scorers on a given student's scores is very small for the KCCT.

Coefficient alpha is not computed for Writing Portfolio scores since each portfolio receives only one holistic score. Chapter 9 provides excellent information on writing portfolio scoring consistency. Briefly, in 2000, 106 schools were identified for writing portfolio audits. Fifty-six schools were selected at random while the remainder were purposefully selected. For the Random group 73.32% of the locally assigned scores were confirmed and for the Purposeful group, 73.76% of the portfolio scores were confirmed by one or more audit readers. For both Random and Purposeful groups, over 99 percent of rescoring were within an adjacent category of the original teacher holistic writing portfolio score.

Student Classifications Accuracy

Since scores include both true achievement and measurement error, and since true *student performance levels* (N/A/P/D) cannot be known, it is necessary to use estimations of (1) the probability that true *student performance levels* is in the same *student performance levels* category (Novice, Apprentice, Proficient, or Distinguished) as the given *student performance levels* level, and (2) the probability that the true *student performance levels* is in another category (measurement error). Thus, for any given standardized test there will be some students that will be misclassified. Hoffman & Wise (2000) use Bayes' Theorem to calculate observed score accuracy. Hoffman & Wise[1] provide a means to answer the often asked student classification accuracy question; "What is the likelihood that the student's unknown true classification is the same as his observed classification?"

Using this approach Hoffman & Wise[2] provide tables for showing the expected proportions of students' true classifications given their assigned classification for each Kentucky Core Content Test (KCCT) grade and content area given in both 1999 and 2000. These tables reveal that extreme student classifications tend to show greater misclassifications. The classic "regression to the mean" may be evident [???]. Thus, the KCCT is more accurate in the middle of the 325 to 800 scale score distribution where more students are found and less accurate at the tails of this distribution where fewer students are found.

---

[1] Hoffman, R. G., & Wise, L. L. (2000). *Establishing the Reliability of Student Level Classifications: The accuracy of observed classifications*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, April, 2000.

[2] Hoffman, R. G., & Wise, L. L. (1999). *Establishing the Reliability of Student Level Classifications: Analytic Plan and Demonstration.*(FR-WATSD-99-34). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., & Wise, L. L. (2000). *The Accuracy Of Students' Novice, Apprentice, Proficient, And Distinguished Classification Of The Kentucky Core Content Test*. (FR-WATSD-00-25). Alexandria, VA: Human Resources Research Organization. CAPS ital

Table 11-2
Student Classification Accuracy
By School Level and Subject

Level	Subject	Year	
		1999[1]	2000[2]
Elementary	Reading	86.09%	87.21%
	Mathematics	78.36%	77.95%
	Science	88.48%	90.37%
	Social Studies	84.75%	85.07%
	Arts & Humanities	79.09%	78.42%
	Practical Living & Vocational Studies	75.26%	74.85%
Middle	Reading	92.58%	93.00%
	Math	75.90%	76.41%
	Science	88.01%	88.09%
	Social Studies	86.63%	86.03%
	Arts & Humanities	78.18%	76.76%
	Practical Living & Vocational Studies	77.19%	78.81%
High	Reading	84.97%	84.66%
	Mathematics	75.68%	76.92%
	Science	90.26%	89.55%
	Social Studies	83.41%	82.83%
	Arts & Humanities	81.82%	81.17%
	Practical Living & Vocational Studies	75.45%	75.85%

[1.] See first citation in footnote 3.
[2.] See second citation in footnote 3

Nine of the 18 grade level/subject area tests have all students classified within one category of their received classification. Eight more have in excess of 99% within two categories. Grade 5 Arts and Humanities is at 98.78% within two categories.

Hoffman & Wise note that if you compare expected and assigned percentages across all grades and subjects that the median difference is about 1%. Thus, they verify that the distribution of expected and observed student classification distributions matches with about 99% accuracy. This overall distribution accuracy is extremely important because student classification data provides 95.0% of the school's accountability index at the elementary level, and 90.0% of the index at the middle and high school levels. Using these data the overall classification accuracy can be computed. Table 11-2 above provides an across year summary of the overall student classification accuracy of the KCCT by grade and subject for 1999 and 2000.

Looking across subjects and years in Table 11-2, the KCCT student performance level classification accuracy for the four classifications (Novice, Apprentice, Proficient and Distinguished) ranges from 74.85 to 93.0, with a median student classification accuracy of 82.33%. This classification accuracy compares favorably with other studies of classification

accuracy. California's CLAS assessment had an exact classification accuracy of 51.72% for their six level performance classification system (Rogosa, 1994). The ACT's Work Keys, five level classification system, showed an exact classification accuracy in the 70% range (Lee, Hanson, and Brennan, 2000)[3]. Hoffman & Wise state that "Given these examples, the Kentucky Core Content Test appears to have classification accuracy statistics that are similar to other educational proficiency assessments."

School Classifications Accuracy

Much like student classification, school classification will not be perfect since school classification is based on student classifications within each school. Hoffman and Wise (2001)[4] provide an overview of the interim accountability system's schools classification accuracy. This interim cycle was used to bridge between the old KIRIS accountability system and the new CATS accountability system. Both KIRIS and CATS have seven assessments (Reading, Mathematics, Science, Social Studies, Arts & Humanities, Practical Living & Vocational Studies, and Writing[5]) given at each school level. However, because of the changes[6] in test structure between KIRIS and KCCT an interim accountably period is necessary to transition between KIRIS and CATS. A regression approach is used (see Chapter 13 for full details) to provide a statistical prediction of how much each school should increase their overall performance for the interim accountability cycle. One standard error of prediction[7] below the regression line (goal line in accountability terms) is provided as a means to reduce errors in school classifications. Schools that are greater than one standard error below their predicted score are classified as schools needing assistance.

Hoffman and Wise established a five-step process to test school classification accuracy in the Interim Accountability Cycle. These steps are:

1. Conduct separate analysis for the standard configuration of elementary middle and high schools.
2. Conduct analyses on three representative school sizes (upper, middle and lower thirds).
3. Conduct generalizability analysis to find the standard error of measurement and reliabilities for each of eight assessments for each of the three representative school sizes. Use this information to compute for each school level and representative school size their standard error of measurement, reliability and accuracy of classification.
4. Synthesize base year (1997 & 1998) standard error and reliabilities.

---

[3] Lee, Hanson, and Brennan (2000), Young and Yoon (1998) and Hoffman & Wise (1999) point to the observation that as the number of categories decrease there is an increase in classification accuracy.

[4] Hoffman, R. G., & Wise, L. L. (2001). The accuracy of school classification for the Interim Accountability Cycle of the Kentucky Commonwealth Accountability and Testing System. (M-00003669). Alexandria, VA: Human Resources Research Organization.

[5] There are two types of writing assessments. One is in the form of a writing prompt given on the KCCT. The other type of writing assessment is a Writing Portfolio that is the accumulation of the student's best work over a period of several years.

[6] The removal of common open-response items between test forms, the addition of multiple-choice items, and the decrease from 12 to 6 forms of the test (for Reading, Mathematics, Science and Social Studies) were some of the more prominent differences between KIRIS and KCCT assessments. The limitation of length of response to constructed response items to one page on the KCCT may have been the most significant change from the school perspective. There was no limit on the KIRIS.

[7] The difference of actual vs. predicted regression school performance as a function of chance alone.

5. Select a conservative reliability estimate for the Non-Academic scores (because there was no known method for estimating reliabilities for these type of data).

Hoffman and Wise provide two versions of reliability, standard error of measurement and classification accuracy information based on two varying assumptions. The first set of test reliabilities, which range from a low of .965 to a high of .980 are for only the combined 1999 and 2000 indices. Standard errors of measurement range from a high of 1.47 index points for small (24 student) elementary schools to a low of 0.52 index points for large high schools (240 students). The standard error of measurement increases as school size decreases. This standard error of measurement and school size relationship is consistent across all three levels of schools (elementary, middle and high school). School classification accuracy ranges between a low of 83.7% and a high of 94.4%. The above analysis assumes that the only measurement error is in the 1999 & 2000 assessment indexes.

An additional analysis is also provided that evaluates the differences between predicted and observed performance. In these analyses some measurement error is assumed in all four years of the interim assessment cycle. Reliabilities are slightly lower and range from a low of .727 to a high of .934 for the combined 1999 and 2000 indices. There is a tendency to have the lower reliabilities at the smaller school sizes[8] in each of the three school levels. The standard error of measurement follows the same general pattern found for reliabilities. Standard errors of measurement range from a high of 2.22 index points for small (24 student) elementary schools to a low of 0.73 index points for large middle schools (240 students). Typically, the standard error of measurement increases as school size decreases. This standard error of measurement and school size relationship is consistent across all three levels of schools (elementary, middle and high school)[9]. School classification accuracy ranges between a low of 73.7% and a high of 89.6%.

Table 11-3

[8] There is a minor reversal for the middle schools for the small (36 student) and the medium (120 student) school with reliabilities at .811 and .800 respectively.

[9] Here again, there is a minor reversal for the middle schools for the small (36 student) and the medium (120 student) school with standard error of measurement at 1.22 and 1.25 index pointes respectively.

School Classification Accuracy
By School Level and Size

Level	Size	Classification Accuracy
Elementary	Small (24)	75.7%
	Medium (60)	87.7%
	Large (96)	89.6%
Middle	Small (36)	82.6%
	Medium (120)	83.7%
	Large (240)	88.1%
High	Small (60)	73.7%
	Medium (168)	81.9%
	Large (240)	84.9%

There are three major classifications that a school can obtain, Meets Goal, Maintaining, and Needs Assistance. As can be seen in Table 11-3, for any given school level or size the exact classification accuracy (that the assigned classification and the school's estimated true classification[10] are the same) is between 73.7% and 89.6%.

However, there is little chance that schools whose true classification was Meets Goal were actually classified as Needs Assistance. Table 11-4 shows that virtually no school[11] that had been misclassified as Needs Assistance when in reality the school should have been classified as Meets Goal. Likewise, there is little chance that schools whose true classification was Needs Assistance were actually classified as Meets Goal. Table 11-5 shows that virtually no school[12] that had been misclassified as Meets Goal when in reality the school should have been classified as Needs Assistance.

Table 11-4
School Misclassification Rate
Meets Goal Schools but Classified Needs Assistance [1]
By School Level and Size

Level	Size	Misclassification
Elementary	Small (24)	0.1%
	Medium (60)	0.0%
	Large (96)	0.0%
Middle	Small (36)	0.0%
	Medium (120)	0.0%
	Large (240)	0.0%
High	Small (60)	0.1%
	Medium (168)	0.0%
	Large (240)	0.0%

1. The estimated percentage of school's whose true classification was Meets Goal but were classified as Needs Assistance (Hoffman and Wise, 2001).

10 True classification assumes that we have a perfectly reliable test and no measurement error in the classification system.

11 Small elementary & high schools had a 0.1% chance of being misclassified as Needs Assistance.

12 Small elementary & high schools had a 0.1% and 0.2% respectively, chance of being misclassified as Meets Goal.

Table 11-5
School Misclassification Rate
Needs Assistance Schools but Classified Meets Goal [1]
By School Level and Size

Level	Size	Misclassification
Elementary	Small (24)	0.1%
	Medium (60)	0.0%
	Large (96)	0.0%
Middle	Small (36)	0.0%
	Medium (120)	0.0%
	Large (240)	0.0%
High	Small (60)	0.2%
	Medium (168)	0.0%
	Large (240)	0.0%

1. The estimated percentage of school's whose true classification was Need Assistance but were classified as Meets Goal (Hoffman and Wise, 2001).

To further clarify any possible misclassification and to assist in achieving the highest possible performance, schools that are classified as needing assistance are provided the opportunity to participate in a Scholastic Audit[13].

Summary

At every point, including scoring reliability, student and school classification the KCCT is found to meet professional standards for test reliability. KCCT reliability compares favorably with other assessments.

[13] Based on the Commonwealth Accountability Testing System assessment results for all Kentucky schools, the lowest one-third of all schools below the assistance line will be audited. Schools in the middle one-third of all schools below the assistance line will have a voluntary review with assistance from Kentucky Department of Education staff. Schools in the upper one-third of all schools below the assistance line will do a self review.

Chapter 12
Reporting to Schools and Districts

Introduction

The Kentucky Department of Education (KDE) notifies schools and districts of their Commonwealth Accountability Testing System performance results on September 15 of each year. This notification includes detailed descriptions of student level and content area scores that lead to a school and district's performance judgment. At the end of a four-year cycle, KDE notifies each school and district of their single performance judgment.

The results for the fourth accountability cycle, a four-year period beginning with the school year 1996–1997 and ending at the conclusion of the 1999–2000 school year, were released to schools and districts on September 15, 2000.

District reports were issued at the same time as school reports and differ from school reports only by being based on all students in the district rather than all students in a school. Thus, district scores at a given accountability grade are not necessarily equal to the weighted sum of the district's school's scores at that accountability grade. The inclusion of scores (or non-cognitive indicator results) from students who attend classes in a special learning environment, or who were not assigned to a reporting school, could alter district results.

School, district, and state-level results were released to the public on September 28, 2000. The embargo period (Sept 15–Sept 28) allowed schools and districts to review their own results and communicate these results to faculty and staff prior to release to the general public.

Individual student results were also provided to schools, in a summary and individual report format. An additional copy of the individual report was provided for distribution to the parents/guardians of students who took the assessment.

Below is a summary list of test result materials sent to schools and districts.

- Individual Student Reports
- Student Listing
- Item Level Report (open-response and multiple-choice items)
- Kentucky Performance Report
- Core Content Report

The remainder of this chapter describes in detail the final Cycle 4 Commonwealth Accountability Testing System reports issued following the 1999–2000 school year.

Individual Student Reports

Two copies of each student's Individual Student Report (Appendix 12-1) were sent to schools: one for students' parents or guardians and one for school use. These reports presented each student's principal performance level (Novice, Apprentice, Proficient, or Distinguished) in as many as three subject areas (reading, mathematics, science, social studies, and writing portfolio), depending on the grade tested. The individual student report also provided performance information not found in any other report presented to schools. The principal performance levels of Novice and Apprentice were further divided into Low (Novice Non-performance replaces the low category for this level), Middle, and High categories. This additional parsing of both the Novice and Apprentice performance levels provides the student with a more precise idea of where his/her achievement is in relation to the next principal performance level.

The individual student reports depict the percentage of Kentucky students scoring in each of these performance levels for each of the subject areas at the student's grade. Each student's Kentucky percentile rank was given in four subject areas of reading, mathematics, science, and social studies. For each subject area, the numeric percentile rank and a visual representation of error bands are provided.

Schools employed a variety of methods to transmit individual student data to the students' parents/guardians. Some schools simply sent the individual student report to parents/guardians, some enclosed letters explaining results, while others asked parents/guardians to attend conferences, at which time results were explained in detail.

Student Listing

The Student Listing (Appendix 12-2) contains information about all students tested or accountable to a particular school by grade level. There are several student accountability types reported on the student listing. These student accountability situations include:

- Students tested and accountable at this school
- Students tested but accountable at another school
- Students tested at another school but accountable to this school
- Students tested with an Alternate Portfolio
- Students tested but exempt from accountability
- Students not tested but exempt from accountability

This listing reports each student's name and "lithocode" identification number, performance level, and Kentucky achievement percentile in each of four open-response content areas tested (reading, mathematics, science, and social studies) for the grade. If a writing portfolio was completed the performance level was also reported.

Schools used the above information to review individual student achievement as well as to ensure there was an accurate accounting of students for whom the school was accountable.

Scores obtained by students who were exempt from testing according to Department of Education policy are not aggregated into the school's total accountability score. However, these students are presented in the student listing, so the school may identify any inconsistency with their records. Scores obtained by students in other accountability situations (noted above) are also presented, informing the school of where each accountable student was tested and what scores were obtained to compute the school's index.

Item Level Report

Much like the student listing, the Item Level Report (Appendix 12-3) gives each student's name and lithocode identification number. However, unlike the student listing report, the item level report provides detailed information about each student's response to each multiple-choice and open-response question and the on-demand writing prompt. Each student's answers to open-response questions were evaluated on a five point, 0-4 scale. Below is the non-grade or item-specific scoring guide that is used as a framework for grade and item-specific scoring.

Scoring Framework (0-4 scale)

- Blank – The score of "blank" indicates a **non-response**. The student made no attempt to answer the question; the answer space was blank.

- 0 – The score point of "0" indicates that a student's answer demonstrated one of two properties. It can mean the student's answer was totally **incorrect**, or it can mean the student's answer was **off-topic**, i.e., had nothing to do with the question, including irrelevant remarks.

- 1 – The score point of "1" indicates that a student's answer demonstrated a **minimal** understanding of the question. The student's response addressed the question but showed little knowledge about the topic. The student did not develop a complete answer and answered only a small portion of the question.

- 2 – The score point of "2" indicates that a student's answer demonstrated understanding of **some** important components of the question. This understanding was clearly communicated. However, the response also demonstrated some gaps in the student's conceptual understanding of the question.

- 3 – The score point of "3" indicates that a student's answer demonstrated an understanding of **most** of the important components of the question. This understanding was clearly communicated. Moreover, the student's response also demonstrated an understanding of the major concepts even though some minor ideas or details were either overlooked or misunderstood.

- 4 – The score point of "4" indicates that a student's answer demonstrated understanding of **all** of the important components of the question. This understanding was clearly

communicated. The student demonstrated in-depth understanding of the relevant concepts or processes. Where appropriate, the student chose the more efficient or sophisticated process. Where appropriate, the student offered insightful interpretations or extensions (generalizations, applications, and analogies).

Multiple-choice responses are displayed as a "+" for a correct answer, a "–" for an incorrect answer, or a "0" for a blank answer.

The student's performance level for reading, math, science, social studies, and on-demand writing are also indicated.

Kentucky Performance Reports

The Kentucky Performance Report (Appendix 12-4) aggregates student level information into either the school or district level. The report contains the following information:

Introduction
Academic Trend Data
Reading Results
Mathematics Results
Science Results
Social Studies Results
Writing Portfolio
On-Demand Writing
Arts and Humanities and Practical Living/Vocational Studies (PL/VS)
Data Disaggregation
Summary Data
Student Questionnaire Results

Introduction

The Kentucky Performance Report introduction provides the reader with an overview of the contents of the report. It furnishes the background for the various parts of the report with regard to grade-specific content areas. The introduction reviews the expectations that all "schools shall expect a high level of achievement of all students." It also describes the exemptions to that standard in the case of a) foreign exchange students, b) medically exempt students, and c) limited-English learners.

Academic Trend Data

The Academic Trend Data reports the 1999–2000 academic index results for each content area assessed. The students' scores have been aggregated by school (or by district for the district report) to produce this index.

Reading Results

The performance results for the content area of reading are reported in this section. The Reading Trend Data page provides the number and percentage of Novice, Apprentice, Proficient, and Distinguished students for the school or district. Within the reading content area, defined by academic expectations, are skills and strategies that students will need to use as they work on the reading test. Important skills are the students' ability to make sense of a wide variety of materials, including literary texts, informational texts, persuasive texts, and practical reading materials. The mean item scores for all items classified in these four subdomains of reading are reported on the reading subscore page. Also included on this page are the results of specific reading questions asked on the student questionnaire.

Mathematics Results

The performance results for the content area of mathematics are reported in this section. The Math Trend Data page provides the number and percentage of Novice, Apprentice, Proficient, and Distinguished students for the school or district. Within the mathematics content area, defined by academic expectations, is a common core of important mathematics skills that students will need to use as they work on the mathematics test. There are four identified reporting subdomains: number/computation, geometry/measurement, probability/statistics, and algebraic ideas. The report lists the school's/district's number and percentage of students at the Novice, Apprentice, Proficient, and Distinguished levels for the 1999–2000 school year. As with reading, the mean item scores for all items classified in these four subdomains are reported on the math subscore page. Also included on this page are the results of specific math questions asked on the student questionnaire.

Science Results

The performance results for the content area of science are reported in this section. The Science Trend Data page provides the number and percentage of Novice, Apprentice, Proficient, and Distinguished students for the school or district. There are three identified reporting subdomains: life sciences, earth and space sciences, and physical sciences. The school's/district's number and percentage of students at the Novice, Apprentice, Proficient and Distinguished levels for the 1999–2000 school year are reported. As with the other content areas, the mean item scores for all items classified in these three subdomains are reported on the science subscore page. Also included on this page are the results of specific science questions asked on the student questionnaire.

Social Studies Results

The performance results for the content area of social studies are reported in this section. The Social Studies Trend Data page provides the number and percentage of Novice, Apprentice, Proficient, and Distinguished students for the school or district. The social studies content area, defined by academic expectations, is concentrated in the following five reporting subdomains: government/civics, culture/society, economics, geography, and history. The report lists the school or district's number and percentage of students at the Novice, Apprentice, Proficient, and Distinguished levels for the 1999–2000 school year. As with the other content areas, the mean item scores for all items classified in these five subdomains are reported on the social studies subscore page. Also included on this page are the results of specific social studies questions asked on the student questionnaire.

Writing Portfolio

The Writing Portfolio performance results are reported in this section. This page displays the number and percentage of Novice, Apprentice, Proficient, and Distinguished portfolios for the school or district.

On-Demand Writing

Similar to the Writing Portfolio, the On-Demand Writing performance results are reported in this section. This page displays the number and percentage of Novice, Apprentice, Proficient, and Distinguished students for the school or district.

Arts and Humanities and Practical Living/Vocational Studies

The performance results for the content areas of arts and humanities and PL/VS are reported in this section. These pages provide the number and percentage of Novice, Apprentice, Proficient, and Distinguished students for the school or district in each content area.

Data Disaggregation

The data disaggregation results report the number and percentage of Novice, Apprentice, Proficient, and Distinguished responses across all content areas for each grade by the following subgroups:

Gender
Ethnicity
Title 1
Migrant Programs
Limited English Proficiency
Extended School Services
Gifted and Talented Programs
Free and Reduced Lunch Program
Students with Disabilities (with and without accommodations)

The data disaggregation process only considers data scanned from student answer documents. To protect anonymity of respondents, no data are reported if a category includes fewer than ten students. The analyses also include students who are participating in the Alternate Portfolio program.

Also included in this disaggregation is the number and percentage of students in each of these categories who participated in the Commonwealth Accountability Testing System at the district, region, and state level. District and state academic indices for each content area are also provided to the schools. Lastly, data on the number of exemptions (medical, limited-English proficiency, and other) is provided at the school, district, region, and state levels.

Summary Data

These pages provide general summary information comparing a school or district's performance to state averages. The academic index , by subject, is provided with up to four years of comparison. The second portion of this report provides demographic information, showing overall number of students and providing information by number and percent by gender, ethnicity, Title 1 service, migrant, LEP, extended school services, gifted and talented, free and reduced lunch, and disability status. In addition, this report provides the number of students having an Alternate Portfolio or exempted status. This report is produced at both the school and district levels and always includes the state comparison number.

This report assists schools in evaluating the general trend of their performances in comparison to overall state trends. It also allows schools to compare their overall population to the overall state population in the listed demographic fields.

Student Questionnaire Results

All students in the accountability system were administered a questionnaire at the end of open-response testing. Questions administered varied slightly from grade to grade. Both the number and percentage of student responses to each question were reported. Questions included length of Kentucky residency and current school attendance, the amount of after-school activities (homework, nonacademic reading, television), the amount of school course coverage in relation to test coverage, and the amount of participation in various activities (groups, projects, information retrieval, oral reports, free-choice reading, etc.). Additional questions included the amount of useful teacher feedback given on homework, the amount and concordance of part-time work with career goals, absenteeism in a given month, vocational plans, current academic performance in school, and current English and mathematics academic level course enrollment. Questions were tailored for each grade level.

Accountability Report

The purpose of the Accountability portion of the report (Appendix 12-5) is to provide a school or district a single, three-page report that has all of the school's Commonwealth Accountability Testing System accountability information. The report provides the accountability index scores by which schools were evaluated. A school's/district's performance judgment will be made using these scores.

The first two pages of the accountability report capture some of the same elements found in the earlier pages of the Kentucky Performance Report. The percentage of Novice, Apprentice, Proficient, and Distinguished at the school or district level is reported by content area. The content areas reported are reading, mathematics, science, social studies, arts and humanities, practical living and vocational studies, and writing portfolio. If a content area is represented by various types of testing (e.g., open-response and portfolio), these values and their weighted totals are also given.

The third page of the accountability report provides schools/districts with their academic, noncognitive, and accountability indices. There is an academic index for each content area. The

content-area-specific academic index is computed by multiplying the percentage of Novice, Apprentice, Proficient, and Distinguished in that content area by 0, 40, 100, and 140, respectively, and summing the products. The weighted academic indices are combined with a weighted noncognitive index to form the accountability index. The mathematics, reading, science, social studies, and writing indices are weighted 14% each, arts and humanities, and practical living and vocational studies are weighted 7% each, and the noncognitive index is weighted 16%.

For schools also interested in reporting only the academic portion of the accountability index, the total academic index is also provided. This index uses only content area weights used with the accountability index to compute the academic index; no noncognitive information is included. However, since the combined cognitive weights only equal 84%, the sum of the products must be divided by .84 to place the resulting total academic index on the same metric as the accountability index.

To provide a school/district with a performance judgment (Reward, Successful, Successful Year 2, Improving, Improving Category 2, Decline, or Crisis), the school or district's Baseline Index, Improvement Goal, and Combined Growth Index must be computed. The accountability report provides all of these indices. The Baseline index, which is the weighted average of the accountability indices for the first two years of the accountability cycle, is used to calculate the amount of growth required of the school or district during the cycle. This growth is added to the baseline to provide the Improvement Goal. The average Accountability Growth Index, which is the weighted average of the accountability indices for the last two years of the accountability cycle, is then compared to the Improvement Goal to produce a performance judgment. Each accountability report has a tailored message indicating the school and district's performance judgment. These reports are further described in the *Interpretive Guide*, published in September 2000, which accompanied the reports.

Core Content Report

The Core Content Report provides information regarding student performance in the individual subdomains and related sections for each content area. Open response and multiple choice information are reported separately. For open response questions, the number of items, observations, percent of students scoring at each point level, the mean, and the standard error are reported for each subdomain and section. Multiple choice reporting lists the number of items, observations, percent correct, percent incorrect, percent omitted or multed, mean, and standard error for these same subdomains and sections. This information is reported at both the school and district levels, each showing corresponding state percentages. A final column shows the variance between the school or district mean and the state mean.

Schools may use the information provided to determine areas where instruction is strong, as well as those areas needing improvement. This information should help guide schools in their instructional planning and professional development.

Conclusion

Reporting the results of the Kentucky Core Content Tests is the culmination of efforts made over several months, if not years, by teachers, students, district staff, the Department of Education, and the testing contractors. The results determine the accountability indexes of the districts, direct the teaching methods and professional development in the schools, and provide feedback to students of content areas in which they excel or in which they are challenged. Therefore, because of the importance of these results, every detail is attended to by the Department of Education and the testing contractors to ensure that the results are accurate and timely.

Chapter 13
Interim Accountability

Introduction

House Bill 53, passed in the spring of 1998, directed the Kentucky Board of Education to redesign the state's assessment and accountability system. Through an extensive and collaborative process involving the educators and citizens of Kentucky, the Commonwealth Accountability Testing System (CATS) resulted. Many changes were made in the system, which was first administered in the spring of 1999. The changes were made in order to improve the reliability and validity of the test, reduce testing time and make the system more equitable and easier to understand.

Because of the major changes in the system, comparisons between KIRIS and CATS are not appropriate. Words like 'gain,' 'growth,' 'improvement,' or 'decline' are not appropriate ways to describe the difference between 1997/1998 KIRIS scores and the 1999/2000 KCCT results of CATS. Because of this lack of comparability, neither the old nor the new Long-Term Accountability models were appropriate for determining rewards and assistance in the year 2000. The National Technical Advisory Panel for Assessment and Accountability (NTAPAA) advised the Kentucky State Board of Education to use a model similar to what American College Testing (ACT) uses to predict student success in college and what doctors use to predict the average growth of a child. In other words, while KIRIS is not CATS, a school's performance on one can be used to understand its performance on the other and to the average performance of all schools on both measures. A simple linear regression model can be used to accomplish this.

The Linear Regression Model and the Distribution of Errors

The linear regression model for the prediction of a dependent variable, Y, from a predictor, X, assumes that the conditional distributions of Y given X are identical for all values of X. This distribution is also known as the distribution of errors about the model. The distribution for a given value, X_i, is centered about the predicted value, $E(Y|X_i)$, and has variance, σ^2. For purposes of making inferences we add the assumption that the errors are normally distributed. Hence we write the model, including assumptions, for any individual (school in our case) as

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \varepsilon_i \\ \varepsilon_i &\sim IN(0, \sigma^2) \end{aligned} \qquad [1]$$

where α is the intercept of the regression line and β the slope. The second line reads "the error term is Independently Normally distributed with a mean of zero and variance sigma-squared. The square root of the variance is referred to as the standard error of estimation. In terms of the model parameters, the expectation is

$$E(Y|X_i) = \alpha + \beta X_i \qquad [2]$$

and an estimate of it is often referred to as the Predicted value of Y. Note from [1] that the errors can be written as

$$\varepsilon_i = Y_i - (\alpha + \beta X_i) \qquad [3]$$

The Kentucky Accountability Regressions

CTB/McGraw-Hill and Kentucky Department of Education staff applied a regression procedure to the interim accountability baseline data (KIRIS data from spring 1997 and spring 1998) and the first and second year data from the new Kentucky Core Content Test (spring 1999 and spring 2000). The interim accountability model maintained the Kentucky accountability system while transitioning from the old KIRIS testing system to the new CATS system. The procedure was consistent with the general principles and guidance provided by the National Technical Advisory Panel on Assessment and Accountability (NTAPAA). The model worked as expected based upon simulations produced using the last biennia of KIRIS data (see Kentucky Accountability Regressions 1997-98 to 1999 Prediction by Carlson). For the purpose of formulas in the remainder of this chapter, the asterisk (*) is used to represent multiplication.

In these analyses the accountability index from 1997-98 was used as the predictor variable, X, and the 1999-00 index was the dependent variate, Y. Clearly it would be easiest for school personnel to understand the computation of the accountability predictions if a single regression model that would be valid for all Kentucky schools could be fit to the data. Unfortunately the same regression model does not fit elementary (E), middle (M), and high (H) schools (see the data reported below). Rather three models were required so rather than [2] the appropriate expression for school *i* is,

$$E(Y|X_{ji}) = \alpha_j + \beta_j X_{ji} \quad (where\ j = E, M, or\ H) \qquad [4]$$

There are also three distinct error terms. For school *i*

$$\varepsilon_{ji} = Y_i - (\alpha_j + \beta X_{ji}) \quad (where\ j = E, M, or\ H) \qquad [5]$$

and three different error distributions with standard errors

$$\sigma_j \quad (where\ j = E, M, or\ H) \qquad [6]$$

Kentucky Department of Education Regulation 703 KAR 5:060 specifies that the standard error will be used to determining each schools' performance category (Reward, Progressing, Needs Assistance). Hence the estimation of appropriate standard errors is very important.

As mentioned above, to simplify interpretation by schools it would be desirable to use a single regression equation for all three types of school (elementary, middle, and high). Using

simultaneous regressions in several populations and procedures outlined in a previous paper (see Kentucky Regression Models by Carlson), CTB fitted 3 regression models and tested hypotheses about whether the data by type of school could be considered to constitute three samples from a single population. The answer was no. The statistical results are displayed in Table 13-1.

Table 13-1
Results of 3 Simultaneous Regression Models

Model	School Type	Intercept	Slope
Unrestricted Simultaneous Regression Model (USRM)			
	Elementary	13.08921	0.97451
	Middle	12.49737	0.92493
	High	17.10304	0.91372
Standard Error Of Estimation = 4.49		*Squared Multiple Correlation = .73*	
Parallel Simultaneous Regression Model (PSRM)			
	Elementary	14.28334	0.94993 (common slope for all 3)
	Middle	11.38444	
	High	15.28497	
Standard Error Of Estimation=4.49		*Squared Multiple Correlation = .73*	
Identical Simultaneous Regression Model (ISRM)			
	Elementary, Middle & High	11.08496	1.00526
Standard Error Of Estimation = 4.69		*Squared Multiple Correlation = .71*	
Statistical Comparisons of Differences Between Models			
Significance of Difference Between USRM and PSRM		*F = 1.116, not significant at the .05 level*	
Significance of Difference Between PSRM and ISRM		*F = 57.179, significant at the .001 level*	

Strictly speaking the presence of combined schools (discussed below) means that the three "samples" are not independent which has some effect on the validity of the *F* tests. However, the first *F* test is so close to the center of the theoretical *F* distribution (1.00) and the second so

far out in the tail of the distribution (beyond the 99.9th percentile) that it seems unlikely that a different conclusion would be reached were it possible to take into account the correlated errors.

The results support a conclusion of parallel (same slope) regressions in the three populations (lack of significance of the first *F* test). But we cannot conclude that the intercepts are the same (significance of the second *F* test). Accepting the parallel model would not accomplish the goal of a single model for schools of all three types. Hence three separate regression models are fitted. A summary of the regression output for the 1999 simulation is displayed in Table 13-2.

Table 13-2
Regressions by School Type

	Intercept	Slope	R-Square	Standard Error of Estimate
Elementary Schools (N=704)				
	13.0892	.9745	.62	5.2
Middle Schools (N=316)				
	12.4974	.9249	.76	3.5
High Schools (N=228)				
	17.1030	.9137	.82	3.2

Another problem with using a common model can be seen from the data in Table 13-2. When we estimate three separate regressions rather than using the unrestricted simultaneous regression model it is found that the three types of schools differ substantially in standard error of estimate. Basically, this means the points cluster more closely about the middle and high school regression lines than they do about the elementary line.

Using a common standard error estimate to determine the schools' classification as reward, progressing or needs assistance would give unfair advantages to some schools and disadvantages to others. Suppose the common value of 4.49 from the parallel or unrestricted models in Table 13-1 was used. Data reported in Table 13-2 indicate that elementary schools actually have substantially larger variation and high schools substantially less. Assuming independence, tests of the significance of the differences of the three variance estimates lead to the conclusion that they are not the same, at the .01 level. Hence, all other things being equal (no real change in performance level), random year-to-year differences would result in more likelihood of an elementary school falling at or below the predicted value minus one standard error. Concomitantly, all other things being equal there would be less likelihood of random differences resulting in a high school index falling in that range. The conclusion is that using a common standard error would not be appropriate.

Combined and Joined Schools

Certain schools in Kentucky are structured such that their students are a mixture of elementary and middle (i.e., grades K through 8), middle and high school (i.e., grades 7 through 12), or elementary, middle and high school students (i.e., grades K through 12). Such schools are designated "Combined Schools". Also, because some schools do not include both grades assessed at a given level (e.g., grades 4 and 5 at the elementary level) certain sets of schools have been designated as "Joined Schools" for purposes of accountability.

There are two (or three) different regression models that apply to the combined schools. For example, one for the elementary students and one for the middle-school students in combined elementary-middle schools. The joined-combined schools also require more than one regression. For these combined and joined-combined schools the Kentucky Department of Education has determined that the school's predicted accountability index will be a weighted average of the two (three) expected values where the weights are the relative proportions of students. For a school that comprises a mixture of N_E elementary students, N_M middle-school students, and N_H high-school students the weights will be computed as

$$\begin{aligned} w_E &= \frac{N_E}{N_E + N_M + N_H} \\ w_M &= \frac{N_M}{N_E + N_M + N_H} \\ w_H &= \frac{N_M}{N_E + N_M + N_H} \end{aligned} \qquad [7]$$

and the predicted index for the school will be

$$\widetilde{Y}_i = w_E E(Y|X_{Ei}) + w_M E(Y|X_{Mi}) + w_H E(Y|X_{Hi}) \qquad [8]$$

For schools that are a combination across all three levels there will be three weights and three terms in [7] and [8]. For schools that are a combination of only two levels the third weight will be zero so a term will drop out of [8]. The standard error of this predicted value should also be computed using the weights. Assuming independence the appropriate standard error should be computed as

$$\sigma_C = \sqrt{w_E^2\sigma_E^2 + w_M^2\sigma_M^2 + w_H^2\sigma_H^2} \qquad [9]$$

As in the previous formula there will only be two terms for schools that do not include all three levels. This formula, like the significance tests discussed earlier assumes independence. Although we know this assumption does not hold we do not know how severe the dependence is. This fact led to the consideration of several alternative methods of computing standard errors.

Note that, because the weights in [9] are fractional and they are squared, the number under the square root, and hence the standard error estimated by the formula will tend to be smaller than either of the component standard errors. That formula is the appropriate one to use for the standard error computation under the assumption of independence. Given the dependencies in the data, however, and the fact that the formula will yield values for many schools that are smaller than either of the component standard errors, an alternative computational method was desirable. To overcome the aforementioned problems, the decision was made to use standard errors of estimate computed from simultaneous regression models for the appropriate combinations of schools. Under this scheme all combined elementary-middle schools would have one standard error, as would each of the other types of combinations. In summary, for the combined schools, the most defensible procedure was to compute the weighted average and use the standard errors from the simultaneous regressions.

Kentucky Interim Accountability Regression Results

As a result of the regression analyses conducted using baseline data (KIRIS data from spring 1997 and spring 1998 combined) and target data from the new Kentucky Core Content Test (spring 1999 and spring 2000 combined), a procedure was conducted according to a process that allowed for separate equations to be generated for the elementary, middle school and high school levels. Likewise, weighted equations and standard errors for combined and joint schools were determined in the manner prescribed in the previous section. The results of these analyses are presented in Table 13-3.

Table 13-3
Regression Results

Elementary School	
Regression Intercept	14.2648
Regression Slope	0.9666
Standard Error of Estimate	4.8
Middle School	
Regression Intercept	12.9406
Regression Slope	0.9336
Standard Error of Estimate	3.2
High School	
Regression Intercept	18.2899
Regression Slope	0.8982
Standard Error of Estimate	3.0
Combined Elementary-Middle School	
Standard Error of Estimate	4.4
Combined Middle-High School	
Standard Error of Estimate	3.1
Combined Elementary-Middle-High School	
Standard Error of Estimate	4.2

Table 13-4 lists the correlations between the 97/98 and 99/00 combined data sets.

Table13-4
Correlations Between 1997-1998 Biennium KIRIS and 1999-2000 Biennium Kentucky Core Content Test Data

School Level	Correlation (r)	R-Squared
Elementary	.80	.65
Middle	.89	.80
High	.91	.82

The following formulas were applied to determine school classifications in 2000:

Elementary Schools:

- Predicted Score = 14.2648 + .9666 * KIRIS Baseline
- Standard Error of Estimate = 4.8
- *Eligible for Rewards* if 2000 Kentucky Core Content Test Index is equal to or greater than Predicted Score
- *Maintaining* if 2000 Kentucky Core Content Test Index is less than Predicted Score and greater than the value (Predicted Score – 4.8)

P-8 Schools:

- Predicted Score = W_E (14.2648 + .9666 * KIRIS Elementary Baseline) +
  W_M (12.9406 + .9336 *KIRIS Middle Baseline)
  W_E = number of elementary school students/total number of students
  W_M = number of middle school students/total number of students * KIRIS Baseline
- Standard Error of Estimate = 4.4
- *Eligible for Rewards* if 2000 Kentucky Core Content Test Index is equal to or greater than Predicted Score
- *Maintaining* if 2000 Kentucky Core Content Test Index is less than Predicted Score and greater than the value (Predicted Score – 4.4)

Middle Schools:

- Predicted Score = 12.9406 + .9336 * KIRIS Baseline
- Standard Error of Estimate = 3.2
- *Eligible for Rewards* if 2000 Kentucky Core Content Test Index is equal to or greater than Predicted Score
- *Maintaining* if 2000 Kentucky Core Content Test Index is less than Predicted Score and greater than the value (Predicted Score – 3.2)

7-12 Schools:

- Predicted Score = W_M (12.9406 + .9336 * KIRIS Middle Baseline) +
  W_H (18.2899 + .8982 * KIRIS High Baseline)
  W_M = number of middle school students/total number of students
  W_H = number of high school students/total number of students
- Standard Error of Estimate = 3.1
- *Eligible for Rewards* if 2000 Kentucky Core Content Test Index is equal to or greater than Predicted Score
- *Maintaining* if 2000 Kentucky Core Content Test Index is less than Predicted Score and greater than the value (Predicted Score – 3.1)

High Schools:

- Predicted Score = 18.2889 + .8982 * KIRIS Baseline
- Standard Error of Estimate = 3.0
- *Eligible for Rewards* if 2000 Kentucky Core Content Test Index is equal to or greater than Predicted Score
- *Maintaining* if 2000 Kentucky Core Content Test Index is less than Predicted Score and greater than the value (Predicted Score – 3.0)

P-12 Schools:

- Predicted Score = W_E (14.2648 + .9666 * KIRIS Elementary Baseline) +
  W_M (12.9406 + .9336 * KIRIS Middle Baseline)+
  W_H (18.2899 + .8982 * KIRIS High Baseline)
  W_E = number of elementary school students/total number of students
  W_M = number of middle school students/total number of students
  W_H = number of high school students/total number of students * KIRIS Baseline
- Standard Error of Estimate = 4.2
- *Eligible for Rewards* if 2000 Kentucky Core Content Test Index is equal to or greater than Predicted Score
- *Maintaining* if 2000 Kentucky Core Content Test Index is less than Predicted Score and greater than the value (Predicted Score – 4.2)

Classification results from Kentucky's 2000, Interim Accountability Cycle (the system that bridged the old system with CATS) are presented below along with some definitions used in that system.

Interim Accountability Cycle

	Meets Goal (Rewards)		Maintaining (Dropout Not Met)		Maintaining*		Assistance		Assistance (Audit)		
School Level	#	%	#	%	#	%	#	%	#	%	Totals
Elementary	376	31.6	N/A		204	17.2	57	4.8	25	2.1	662
P-8	39	3.3	N/A		46	3.9	2	0.2	7	0.6	94
Middle	99	8.3	N/A		76	6.4	10	0.8	13	1.1	198
7-12	15	1.3	1	0.1	11	0.9	3	0.3	0	0.0	30
High	88	7.4	6	0.5	73	6.1	28	2.4	4	0.3	199
P-12	1	0.1	0	0.0	4	0.3	0	0.0	0	0.0	5
Totals	618	52.0	7	0.6	414	34.8	100	8.4	49	4.1	1188

*One (1) school was designated a "Maintaining" school for the following reason: The school's final accountability index, which is the weighted average of its 1998-1999 and 1999-2000 accountability indices, was equal to or less than its assistance point for the Interim Accountability Cycle, but was greater than 80.

Meets Goal

Each school's final accountability index, which is the weighted average of its 1998-1999 and 1999-2000 accountability indices, meets or exceeds its predicted performance for the Interim Accountability Cycle. These schools have been designated as "Rewards" schools for the Interim Accountability Cycle.

Maintaining/Dropout Not Met
Each school's final accountability index, which is the weighted average of its 1998-1999 and 1999-2000 accountability indices, meets or exceeds its predicted performance for the Interim Accountability Cycle. However, each high school's student dropout rate is not meeting the dropout criteria. These schools have been designated as "Maintaining" schools for the Interim Accountability Cycle.

Maintaining
Each school's final accountability index, which is the weighted average of its 1998-1999 and 1999-2000 accountability indices, is less than its predicted performance and greater than the assistance point for the Interim Accountability Cycle. These schools have been designated as "Maintaining" schools for the Interim Accountability Cycle.

Assistance
Each school's final accountability index, which is the weighted average of its 1998-1999 and 1999-2000 accountability indices, is equal to or less than its assistance point for the Interim Accountability Cycle and is less than or equal to 80. In addition, each school has scored in the top two-thirds (2/3) of the schools classified as "Assistance" based upon their final accountability index. These schools have been designated as "Assistance" schools for the Interim Accountability Cycle and shall develop a school improvement plan and are eligible to apply for Commonwealth School Improvement Funds.

Assistance Audit
Each school's final accountability index, which is the weighted average of its 1998-1999 and 1999-2000 accountability indices, is equal to or less than its assistance point for the Interim Accountability Cycle, and is less than or equal to 80. Each school has scored in the bottom one-third (1/3) of the schools classified as "Assistance" based on their final accountability index. These schools have been designated as "Assistance Audit" schools for the Interim Accountability Cycle and are subject to an interim scholastic audit, shall develop a school improvement plan and are eligible to apply for Commonwealth School Improvement Funds.

Chapter 14
Validity

Introduction

Validity has been described as "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores.[1]" This discussion utilizes three broad, traditional categories of validity evidence necessary to support such inferences: content-related, criterion-related and construct-related. These traditional notions of validity have been supplemented with specific criteria for performance assessments (e.g., Frederiksen & Collins[2], Linn, Baker & Dunbar[3]) as well as the idea that the consequences of using a given test are an important aspect of validity. Consequential validity addresses the issue of whether a test has the effect it is intended to have. The description and uses of consequential validity were proposed and advanced by Samuel Messick[4] of the Educational Testing Service.

Intended Goals of the Kentucky Assessment Program

While the main focus of the present Technical Report is on school years 1996-1997 through 1999-2000, the current chapter reports validity evidence both before this time period (KIRIS) and during this period (CATS). Since the establishment of KERA, the role of the Kentucky assessment program is to promote educational improvement for all children in the state. It does this in three major ways:

1. The assessment program provides goals, standards, and criteria for educational achievement. These goals, standards, and criteria are linked together throughout the assessment program. They include the statement of goals in the KERA legislation, the academic expectations, the core content for assessment, and the links between these and specific items and their scoring guidelines. The assessment program includes operational definitions of success, various academic performance levels, and relative weights for assessment components.

2. The assessment program provides useful information on progress made by schools towards achieving those goals. Although the major informational use of assessment scores is in relation to school accountability as mandated in the KERA legislation, much of the assessment data also proves useful in monitoring achievement and progress of individual students, the state as a whole, and various groups of students within the state.

3. The assessment program provides useful information on potential differential impact of the assessment program within the school, district, region, and state for various subgroups such as gender, ethnic and racial minorities, and children receiving Title I assistance.

---

1 American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington: APA

2 Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

3 Linn, R. L., Baker, E. L. & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

4 Messick, S. (1989). "Validity", in R.L. Linn (ed.), Educational Measurement, Third Edition, MacMillan Publishing Co., 1989.

Content and Construct Validity

Although consequential validity concerns may ultimately prove more important than issues of technical quality, it remains very important to examine the KCCT score validity from a traditional psychometric perspective. Thus, content validity information is reviewed below. Construct-related validity evidence is discussed based on the relationships of KCCT tests with each other, with scores from other testing programs, and with qualitative criteria for judging schools. Intermingled with traditional notions of validity in this analysis are more recently proposed criteria for evaluating performance assessments: systematic validity, directness and transparency[5]; and fairness, transfer, generalizability, cognitive complexity, content quality, content coverage, meaningfulness, cost, efficiency, and consequences[6].

Content-Related Validity Evidence

A previous chapter of this manual describes how the components of the KCCT assessment are derived from Kentucky's six Learner Goals and the 57 Academic Expectations, using advisory committees of Kentucky teachers to make those outcomes and expectations operational through test items, and to make choices about what the tests should contain. Many tables in Chapter 2 summarize the academic expectations as well as the distribution of the core content measured by the items in the assessment, and there is little to add to the extensive treatment of this material. In short, there is substantive content-related validity evidence in the process by which the KCCT assessments are constructed.

While test development information serves as the primary source of content-related validity evidence, examining the KCCT tests in terms of the novel content-relevant criteria noted above provides a potential source of additional evidence. Cognitive complexity, content quality, and content coverage can serve as criteria by which to evaluate performance assessments. Because there exist no established standards for these criteria (as noted by Linn, Baker and Dunbar), any detailed consideration of them probably requires discourse substantiated by expert judgment in the form of task analysis. The presence of teacher and other content area specialist judgment in writing and selecting items for the assessments provides a good positive indication of content validity.

Construct-Related Validity Evidence

Because the KCCT testing program assesses student performance in several content areas using a variety of testing methods, it is important to study the pattern of relationships among such content areas and testing methods. One method for studying patterns of relationships to provide evidence supporting the inferences made from test scores is the multi-trait, multi-method matrix (see *KIRIS Accountability Cycle 1 Technical Manual*). Another method for studying patterns of relationships among varying types of test or item scores is factor analysis. To provide evidence for the construct validity of KIRIS open-response item scores, factor analysis was performed on

5Frederiksen & Collins.
6Linn, Baker & Dunbar.

data obtained from open-response scores from the 1993 through 1996 KIRIS testing program (see *KIRIS Accountability Cycle 2 Technical Manual* for a detailed description of this analysis).

Concurrent-Related Validity Evidence

A measure of validity is how well a test correlates with accepted measures of the same or similar constructs. To the extent that few or no other "primarily performance-based" assessments exist for comparison with KCCT, options are limited for demonstrating concurrent validity--although many traditional measures have been enhanced to include constructed-response supplements in conjunction with multiple-choice or selected-response items. The best one can do is to compare the performance of students on the KCCT to accepted or "traditional" tests of academic achievement, despite the fact that they will not assess exactly the same construct as the KCCT. Correlations with the KCCT should be moderate since most norm-referenced tests have many content area requirements in common with the KCCT; however, KCCT has additional higher order thinking requirements for test items that many norm-referenced multiple-choice tests do not possess. Concurrent validity can be assessed through correlational study using different units of analyses (1) the student, (2) the school, and (3) the state.

1. **Student-Level Relationships.** In addition to providing concurrent validity evidence, a good reason for comparing the KCCT to traditional forms of assessment is that those traditional measures are still in use. Tens of thousands of Kentucky high school students take the American College Test (ACT) each year for college admissions. If the KCCT proved to be uncorrelated with that measure, it would place students, parents, and teachers in the uncomfortable position of having to choose the test on which they would like to focus their attention. Hoffman[7] correlated high school juniors' and seniors' ACT scaled scores and their theta scores on KIRIS. Student scores from the years 1994, 1995, and 1996 comprised the data. The sample of students who took the ACT had higher open-response scores than Kentucky students in general. This difference indicated that the results of this study might be generalized to only the upper portion of the distribution of high school juniors and seniors. The strongest relationships were: KIRIS Reading and ACT English scores, r = .56; KIRIS Reading and ACT Reading, r = .52; KIRIS Reading and ACT Composite, r = .56; KIRIS Math and ACT Math, r = .72; KIRIS Math and ACT Composite, r = .70; KIRIS Science and ACT Science r = .57; KIRIS Science and ACT Composite, r = .62. For this group of high-school students (N=51,967), there were moderate to high, positive, linear correlations between these scores. The relationships were stronger between mathematics scores, however, the author reported no test reliabilities and it may be that the higher correlations between mathematics tests were due to higher reliabilities of the mathematics assessments.

   Wise[8] described initial results of efforts to link scores from the Armed Services Vocational Aptitude Battery (ASVAB) and KIRIS for Kentucky high school students for the years 1993 through 1996. The number of students matched by year by grade ranged from 3,567 to 16,314. Total students matched for all years were 64,278. Using data for the years 1994, 1995, and 1996, the student-level score (scaled and theta scores for the respective tests)

7 Hoffman, R. G. (1998) Relationships Among KIRIS Open-Response Assessments, ACT Scores, and Students' Self-Reported High School Grades. Radcliff, KY: Hunan Resources Research Organization.
8 Wise, L. L. (1997) Merging ASVAB and KIRIS On-Demand Scores: Report of Preliminary Results. Ratcliff, KY: Human Resources Research Organization.

correlations for reading ranged from r =. 51 to r = .56. The correlations for the KIRIS math scores for these years were considerably higher, the range was r = .63 to r = .73. By contrast, the correlations for science were somewhat lower, ranging from .42 to .58. The correlations for the 1993 KIRIS assessments with ASVAB were somewhat lower than in the other years compared. The author suggested that this might have been due to somewhat lower reliabilities for KIRIS for that year. Other ASVAB tests, designed to measure nonacademic areas of achievement, for example, Auto and Shop Information, did not match KIRIS subject matter. These positive, linear, moderate relationships indicate that KIRIS measures constructs similar to those measured by the content-matching subscales of ASVAB.

2. **School-Level Relationships.** Considering that the KCCT scores are used for school accountability, it may be argued that a high degree of relationship between the KCCT and other test scores obtained by schools is even more essential evidence of concurrent validity than correlations among student level scores. Obtaining evidence of this kind is problematic insofar as few other tests and fewer content areas are administered to all students in a school, in contrast to the KCCT, which has seven different content areas and is given to about 99% of all students in most years. (For the KCCT, the remaining students are exempted typically for medical or language reasons, or participate in school accountability through the alternate portfolio program.) Of these other tests, very few are administered to a representative or even approximately random sample of students at participating schools, further diminishing the meaningfulness of school-level comparisons.

   Hoffman does provide some insight about the relationship of high school ACT scores and KIRIS scoring. However, schools' means are calculated using only the ACT-taking population of students. "Schools with high scores on ACT also have high open-response scores. At the school level of analysis, GPA is not related to either open-response or to ACT. This is presumably due to differences in grading standards between schools. When gains in schools means are calculated using only the ACT-taking population of students, schools whose ACT-taking students are gaining on any one of the assessments tend to be gaining on all of the assessments, including open-response, ACT, and GPA. This result is obtained in spite of the typically unstable nature of correlational examinations of score gains" [9]

   Because the core content for assessment did not change much in the transition from KIRIS to the Commonwealth Accountability Testing System (CATS), one would expect the above relationships to be similar to those that will result from similar studies using KCCT data. Many of these analysis are planned and underway.

3. **State-Level Relationships.** Considering that the KCCT is administered only in Kentucky, there are no other states with which to compare student performance on the KCCT. However, it is possible to compare changes in state-level scores on the KCCT over time with state-level changes over time on other measures. The challenge in making such a comparison is that, relative to the first year of testing, some improvement in the KCCT scores is likely to occur as a result of directing school curricula toward the test and familiarizing students with responding to open-response questions in general. Initial gains from the 1992 baseline were unlikely to generalize to other tests, but were a predictable,

9 Hoffman, R. G. (1998) Relationships Among KIRIS Open-Response Assessments, ACT Scores, and Students' Self-Reported High School Grades. Radcliff, KY: Hunan Resources Research Organization.

initial result of implementing a high-stakes testing program. To the extent that this effect has been observed with multiple-choice tests[10] used in a high-stakes setting, a finding that initial KIRIS gains did not generalize to other tests was not evidence against score validity, but rather an indication that caution must used in interpreting score gains relative to the first few years of high-stakes testing.

The best available comparison in this regard is the National Assessment of Educational Progress (NAEP). During 1999 and 2000, Kentucky participated in the NAEP program. NAEP is a standards-based assessment that is administered to a national sample. NAEP is also administered at the state level, to a different sample of students. The state NAEP assessments are not aggregated to obtain the national results. Kentucky has participated in all of the assessments since NAEP began state testing in 1990. The data in Table 14-1 summarizes school participation rates, numbers of schools, student participation rates, and the total number of students assessed for all state NAEP administrations in Kentucky.

For each state administration, NAEP selects a sample of approximately 100 schools and approximately 2,500 students per subject per grade. The tests are administered at grade 4 and/or grade 8. The state sample is stratified by characteristics such as urban/rural, percentage of minority students, median household income, education of residents over 25, and other demographic data. Some characteristics are not used on some state tests, or during certain years. Within the strata, the schools are chosen randomly, and within the school, approximately 30 students per subject per grade were chosen randomly. In 1998, 2,442 students participated in the NAEP Reading test in grade 4 while 2,282 students took the NAEP Reading test in grade 8. All these students were public school students. Results are not reported at the district, school, or student level. However, state NAEP results are reported when participation rate requirements are met. More than 70 percent of the initial sample must participate for state NAEP reporting purposes. Notations are made if the initial sample participation falls below 85 percent, *and* if the school participation level falls below 90 percent after substitutions.

The United States Department of Education administers the NAEP through the National Center for Educational Statistics (NCES) and its contractors. NCES has primary responsibility for overseeing planning, development, production, testing, sampling, training, scoring, analyzing and reporting. The Educational Testing Service (ETS) performs item development and field-testing. National Computer Systems distributes and processes materials. Westat manages the field administration of the assessment.

10 See, for example, Linn, R. L. (1995). Assessment-based reform: Challenges to educational measurement. Educational Testing Service: Princeton, NJ.

Table 14-1
National Assessment of Educational Progress
Kentucky Participation Rates in the State NAEP

	Weighted School Participation Rate (%)		Number of Schools Participating	Weighted Student Participation Rate (%)	Total Number Of Students Assessed
	Before Substitutes	After Substitutes			
READING					
Grade 4					
1992	94	97	119	96	2752
1994	88	96	101	97	2758
1998	90	92	99	96	2442
Grade 8					
1998	87	87	91	93	2282
MATHEMATICS [1]					
Grade 4					
1992	93	96	118	96	2703
1996	88	96	102	95	2579
2000	92	94	104	95	2275
2000[2]	92	94	104	95	2335
Grade 8					
1990	100	100	104	95	2680
1992	96	98	104	96	2756
1996	88	92	101	94	2461
2000	94	95	97	94	2294
2000[2]	94	95	97	94	2363
SCIENCE [1]					
Grade 4					
2000	92	94	105	95	2248
2000[2]	92	94	105	95	2311
Grade 8					
1996	87	92	100	94	2459
2000	94	95	96	94	2303
2000[2]	94	95	96	94	2383
WRITING					
Grade 8					
1998	87	87	89	93	2341

1. "…two different sets of NAEP results {are} based on the split-sample design:
   - those that reflect the performance of regular and special-needs students when accommodations were not permitted, and
   - those that reflect the performance of regular and special-needs students— those who required and were given accommodations (such as extended time, small group administration, Spanish- English bilingual booklets, etc.) and those who could be tested without accommodations—when accommodations were permitted." (NAEP Mathematics Report Card, p6)
2. With accommodation

Table 14-2 indicates the percentages of Kentucky students who fell into NAEP's categories of basic, proficient and advanced. The Table also provides data for comparison with the Southeast region and the nation. Only the percent below and percent at or above basic add to 100 percent. The other two columns are included in the above basic percentage. This Table demonstrates the tests that were administered in Kentucky, the percentage of students below basic has declined, and the number at or above proficient has increased in every case.

The results of the 1998 NAEP 4th grade reading test revealed:

- Kentucky is one of only three states to make statistically significant gains from 1992 to 1998 and from 1994 to 1998. (Connecticut and Colorado were the others.)
- Kentucky, compared to 1992, increased five points while the nation had no increase and the southeast dropped one point.
- Kentucky 4th grade readers started out two points below the national average in 1992, equaled the national average in 1994, and moved three points above it in 1998.
- For 4th graders scoring proficient or better, from 1992 to 1998, Kentucky jumped six percentage points while the nation increased two and the southeast one.
- The percentage of 4th graders reading at the lowest level by national standards decreased five percentage points in Kentucky while declining only one percentage point for the nation, and increasing one point for the southeast.
- Kentucky's 8th graders outscored both the nation and the southeast in reading.

The demographic data for 4th grade NAEP results show some important gains but also highlight significant challenges. For example, Kentucky's male students gained 10 points from 1994 to 1998, while females gained 3, narrowing the gender gap, 4th grade girls in Kentucky are now outperforming boys by only 4 points. But the 25-point gap between black and white scores in Kentucky, while narrower than the national racial gap of 32, remained constant from 1994 to 1998, with African-American students and white students each gaining six points.

Table 14-2
National Assessment of Educational Progress
Comparison with Southeast and Nation

	Scale Score	Percent Below Basic	Percent At Or Above Basic	Percent At Or Above Proficient	Percent At Or Above Advanced
READING					
Grade 4					
1992					
Kentucky	213	42	58	23	3
Region	211	45	55	22	4
Nation	215	40	60	27	6
1994					
Kentucky	212	44	56	26	6
Region	208	47	53	23	6
Nation	212	42	58	28	7
1998					
Kentucky	218	37	63	29	6
Region	210	46	54	23	5
Nation	215	39	61	29	6
Grade 8					
1998					
Kentucky	262	26	74	29	2
Region	258	32	68	25	2
Nation	261	28	72	31	2
MATHEMATICS					
Grade 4					
1992					
Kentucky	215	49	51	13	1
Region	210	54	46	11	1
Nation	219	43	57	17	2
1996					
Kentucky	220	40	60	16	1
Region	216	47	53	14	2
Nation	222	38	62	20	2
2000 - no accommodations					
Kentucky	221	40	60	17	1
Region	220	41	59	19	1
Nation	226	33	67	25	2
2000 - with accommodations					
Kentucky	219	41	59	17	1
Region	221	41	59	19	2
Nation	225	35	65	23	2
Grade 8					
1990					
Kentucky	257	57	43	10	1
Region	254	58	42	12	1
Nation	262	49	51	15	2
1992					
Kentucky	262	49	51	14	2
Region	259	53	47	13	1
Nation	267	44	56	20	3

Table 14-2 (continued)
National Assessment of Educational Progress
Comparison with Southeast and Nation

	Scale Score	Percent Below Basic	Percent At Or Above Basic	Percent At Or Above Proficient	Percent At Or Above Advanced
1996					
Kentucky	267	44	56	16	1
Region	264	46	54	16	2
Nation	271	39	61	23	4
2000 - no accommodations					
Kentucky	272	37	63	21	3
Region	265	46	54	18	3
Nation	274	35	65	26	5
2000 - with accommodations					
Kentucky	270	40	60	20	3
Region	263	47	53	18	3
Nation	273	37	63	26	5
SCIENCE					
Grade 4					
2000 - no accommodations					
Kentucky	152	30	70	29	3
Region	141	44	56	21	2
Nation	148	36	64	28	3
2000 - with accommodations					
Kentucky	152	31	69	28	2
Region	141	44	56	21	2
Nation	147	38	62	27	3
Grade 8					
1996					
Kentucky	147	42	58	23	2
Region	141	49	51	21	1
Nation	148	40	60	27	3
2000 - no accommodations					
Kentucky	152	38	62	29	3
Region	143	48	52	24	3
Nation	149	41	59	30	4
2000 - with accommodations					
Kentucky	150	40	60	28	3
Region	142	49	51	23	3
Nation	149	41	59	30	4
WRITING					
Grade 8					
1998					
Kentucky	146	16	84	21	1
Region	143	19	81	19	1
Nation	148	17	83	24	1

Reading

In 1998, the NAEP reading test was administered to 2,442 4th grade students in 99 randomly selected Kentucky schools and to 2,282 8th graders in 91 randomly selected schools. The NAEP reading tests consist of three kinds of questions: multiple choice, short answer (generally requiring a sentence) and essay (requiring a paragraph or two). They are scored on a 0-500 scale.

Writing

Kentucky 8th graders who participated in the 1998 NAEP writing assessment scored near the national average and above other states in the southeast region. 1998 marked the first time state-level NAEP writing tests were administered to Kentucky 8th graders. The full 1998 NAEP assessment included state-level assessments in reading at grades 4 and 8 and in writing in grade 8. The percentage of Kentucky 8th graders performing at or above the basic achievement level was slightly higher than both the southeast and nation. Only eight states had significantly higher average scale scores than Kentucky: Colorado, Connecticut, Maine, Massachusetts, Oklahoma, Texas, Virginia and Wisconsin. It should be noted that some of the states that have higher average in writing (i.e. Connecticut, Maine and Texas) are states, like Kentucky, that have on-demand writing assessments.

The NAEP writing assessment was given to samples of 8th graders in 35 states in 1998. The NAEP is generally considered to be the only test given in the U.S. that gives valid results that can be compared from state to state.

Mathematics

The NAEP 2000 mathematics assessment was given to samples of 8th graders in 40 states in 2000. Results from the Kentucky 2000 NAEP mathematics assessment show that 8th graders' performance improved significantly from the 1996 results. Eighth-graders' performance moved from an average scale score of 267 to 272, a gain of five points, compared with gains of one point in the southeast and three points in the nation. Kentucky 4th graders' performance, which improved by five points between 1992 and 1996, did not improve significantly in 2000, moving from an average scale score of 220 to 221.

Students' 8th grade NAEP mathematics performance showed an increase of five percentage points at the proficient level, while students in the southeast increased by two percentage points and those in the nation by three percentage points. Even though the percentage of 8th graders who performed at or above proficient (21 percent) was smaller than the national percentage (26 percent), Kentucky's percentage was significantly higher than it was in 1996. The percentage of students moving from below basic to at and above basic and from proficient to advanced also showed significant increases.

Science

NAEP science scores show significant growth at the 8th grade level for Kentucky from 1996 to 2000. In 2000, NAEP tested nearly 500,000 4th and 8th grade students in participating states and jurisdictions in mathematics and science. Nearly 2,300 Kentucky 8th graders participated, and they scored an average of 152 in science. This is the first time Kentucky 8th grade students scored higher than the national average on the NAEP science test. The score also represents a five-point increase since 1996, which NAEP considers significant. Only three states (Kentucky, Vermont and Missouri) have made significant 8th grade science progress from 1996 to 2000 on NAEP.

2000 was the first year 4th graders were tested in the NAEP science component. Kentucky's 4th-graders performed close to the national average and above other southeastern states. Kentucky's scale score for 4th graders was higher than those in 19 states or jurisdictions; not significantly different from those in 13; and lower than those in 11 states. The scale score for Kentucky's 8th graders was higher than those in 16 states or jurisdictions; not significantly different from those in 11; and lower than those in 14. Although Kentucky's African-American students outperformed the nation's African-American students on average in both the 4th and 8th grades in 2000, the NAEP achievement gap at the 8th grade level has actually gotten worse.

Caveat

It should be noted that there are methodological issues related to scaling in making comparisons across measures. Not only is each test built to its own specifications, but also each measure has its own scale. As long as each measure provides an indication of whether changes over time are statistically significant[11], it is possible to compare trends broadly. Comparing the magnitude of changes on one measure with magnitude of changes on another is more complicated, especially when multiple sets of scores are available for one or the other of the measures (such as theta and standards-based –Novice, Apprentice, Proficient, Distinguished – scores on the KCCT).

Consequential Validity

A vast potential source of validity evidence to support or refute the inference that accountability gain scores reflect improvements in school performance is schools themselves. The primary challenge associated with taking advantage of this rich source of information is that it is logistically difficult (and therefore expensive) to gather meaningful data on schools. A lesser challenge (and, some would argue, a potential advantage) is that information of this nature does not necessarily lend itself readily to quantification, so that results must be considered mostly in qualitative terms.

A case study of 16 schools conducted by Kelley[12], a senior research associate at the Wisconsin Center for Educational Research, provided important initial evidence for criterion-related

11 The use of effect size (Cohen, 1998, 1994) rather than statistically significance, could be used to analyze and interpret differences between two groups. Effect size could be an effective way of measuring the effectiveness of educational interventions.
12 Kelly, C. and Protisik, J. (1997). Risk and reward: Perspectives on the implementation of Kentucky's school-based performance award program. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

validity. Kelley found that successful schools had taken specific actions to achieve success, including analyzing test results to identify weaknesses, setting goals, changing curriculum and using professional development effectively. By contrast, low-success schools had not changed their curriculum and had not used professional development to learn about the new learning goals. These findings are encouraging, but more data are necessary.

The importance of the consequential validity of the test becomes obvious when one thinks of the alternatives. The KCCT exists because the Kentucky legislature believed that the KCCT would increase the likelihood of KERA's success by supporting (indeed, driving) changes in the classrooms of Kentucky. Thus, all decisions related to the KCCT ultimately have to be considered in light of the question, "Will this change lead to better instruction and more real achievement?" If not, the change becomes less justified.

Evidence and Interpretation of Consequential Validity

The KERA legislation and subsequent programs, including the assessment program, have engendered much discussion and activity. There is some formal research available on the impact of the assessment program on classroom practice, teacher development, or support of educational reform in Kentucky. In the available research it is often difficult to separate effects of the assessment program from other aspects of educational reform.

It should be noted that the discussion in this chapter of the consequential validity of the Kentucky assessment program paints broad strokes, which may not apply to every classroom in every school. Change is taking place at different speeds and in different forms throughout the Commonwealth, often for different combinations of reasons. In addition, the assessment program, and people's understanding of it, has changed over time. Continuing research will be required to provide more complete results of the consequential validity of the Kentucky assessment program, as well as to keep research findings up to date.

Consequences: Provides Goals, Standards, and Criteria for Instruction and Curriculum

Evidence continues to accrue regarding the effects of the KCCT and its predecessor on instructional practice, teachers' professional development, and support for educational reform. Evidence presented in the *KIRIS Accountability Cycle I Technical Manual* addressed impact on instructional practice, professional development, and educational reform, citing several studies.[13]

13 Appalachian Educational Laboratory. (Dec., 1994). Instruction and assessment in accountable and non-accountable grades, Notes from The Field, 4(1), 1-2.; Pankratz, R., Ochs, D. et al. (April, 1995). Configuration maps: Results from Kentucky. Papers presented at the annual meeting of the American Educational Research Association, San Francisco, CA; Policy Studies Associates, Inc. (1994). Third-year evaluation of the nine-site program initiative. (A report to the U.S. Department of Education.) Washington, DC: Author; McCollum, H. et al. (August, 1994). Portfolio assessment in mathematics: Views from the classroom. Annual report. Washington, DC: Policy Studies Associates, Inc.;Roberts, R. & Kay, S. (September, 1993). Kentuckians' expectations of children's learning: The significance for reform. A public report prepared for the Prichard Committee for Academic Excellence and the Partnership of Kentucky School Reform, Lexington, KY: Roberts & Kay, Inc. (Available from the Prichard Committee for Academic Excellence, P.O. Box 1658, Lexington, KY 40592-9980.); Winograd, P., Jones, D., & Perkins, F. (submitted). The politics of alternative assessment: Lessons from Kentucky. (Manuscript obtained from first author. ; Laguarda, K. G., Breckenridge, J. S., Hightower, A.M., & Adelman, N. E. (September, 1994). Assessment programs in the statewide systemic initiatives (SSI) International, primary contractor.) Prepared under contract for the National Science Foundation, SRI International, primary contractor. Washington, DC: Policy Studies Associates, Inc.

Taken as a whole, those early studies suggested that KIRIS had an impact on instructional practice.

After the completion of the *KIRIS Accountability Cycle I Technical Manual*, several other studies took place that addressed consequential validity in the context of curriculum and instruction. One major study by RAND[14] was cited in the *KIRIS Accountability Cycle 2 Technical Manual.* This survey and interview-based study found that "about 40 percent of the teachers reported that the open-response items and portfolios have had a great deal of positive effect" on instruction, with only half as many teachers endorsing this view of performance events, and almost none considering multiple-choice items to have had such a positive effect. It should be noted that the study examined a broad range of perceived effects, including perceived negative impact on instruction, and provides a more extensive discussion of its findings than can be afforded in this chapter.

Consequences: Provides Information on Status and Progress

The KCCT provides information to schools and districts in several forms. These are described more fully in the chapter on Reporting in this *Technical Report*. Essentially, KCCT reports include scores on each subject matter area and a nonacademic index for schools and districts each year. The Department and its contractors produce reports that summarize each school's performance in terms of a two-year baseline and two subsequent years of the accountability cycle, and summarize the school's status in relation to rewards and assistance. Student reports are produced each year for subject matter areas, and the writing portfolio. The student reports are sent by the contractor to schools, where they are distributed to parents by different systems determined by the school. In addition, each year KDE and its contractors produce a summary report for the state, by region, by gender, by ethnic group, and disabilities.

Schools, districts, and classroom teachers report using the score reports in a variety of ways consistent with the intent of the KCCT. The most common use is in broad program review to mark progress over a year or two, and to focus resources for instructional program improvement. Analysis of KCCT scores comprises an essential part of every school's annual Comprehensive School Improvement Plan (formerly Consolidated Plan).

However, KCCT scores have been used by schools and teachers for other purposes. There are occasional requests to have the KCCT provide information in addition to the school accountability function it was originally designed to provide. There have been calls for additional information in the other traditional evaluation areas:[15] Some examples are listed below:

1. Individual student achievement status for use on report cards;
2. Individual student comparative status for college admissions;
3. School comparisons (the news media routinely convert reports into "rankings" that facilitate comparisons between schools and districts);

14 Koretz, D. M., Barron, S.,Mitchell, K. J., & Stecher, B. M. Perceived effects of the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: RAND.

15 For example, see the report done by The Evaluation Center, Western Michigan University, (January, 1995). *An independent evaluation of the Kentucky Instructional Results Information System (KIRIS).*

4. Student diagnostic information for monitoring student progress and informing instructional changes by classroom teachers;
5. Instructional program evaluation (e.g., to monitor and improve instructional programs, school curricula, and inform teacher assignments and professional development).

While the KCCT provides results that address these areas to some extent, most would require substantial changes in KCCT design and/or operation. Some of these uses are possible; and some are possible but somewhat incompatible with the intended uses of KCCT results. For example, using student KCCT results as the sole basis for school report card grades is viewed by KDE as inappropriate. Providing diagnostic information for individual students would require not only a complete revamping of the KCCT test but also a much more rapid feedback than is possible. It should be remembered that KCCT is a school accountability assessment and that it is not designed as a student diagnostic instrument. KDE believes that such diagnosis is more appropriately undertaken by classroom teachers using classroom assessments other than or, at least, in addition to, the KCCT assessments, and is more appropriately undertaken earlier in the school year (i.e., much earlier than April which is near the end of the school).

Another important caution should be noted because of the redesign the state's assessment and accountability system during the accountability cycle. Through a broad and collaborative process involving educators and citizens of Kentucky, the Commonwealth Accountability Testing System (CATS) resulted. Many changes were made in the system, which was first administered in the spring of 1999. The changes were made in order to improve the reliability and validity of the test, reduce testing time and make the system fairer and easier to understand. These changes include, but are not limited to:

- Distributing the test components for the high school from primarily the junior year to across three grade levels: Reading and Practical Living/Vocational Studies in grade 10; Mathematics, Science, Social Studies and Arts and Humanities in grade 11 and Writing On Demand and Writing Portfolios in grade 12
- Reducing the contents of the Writing Portfolio in each accountability year – grade 4 from six to four pieces and grades 7 and 12 from six to five pieces; also creating a regulation that directs instructional use, editing and scoring of the portfolio
- Limiting student answers on the open-response questions to the space provided – one 8 ½" x 11" sheet for each open-response question
- Counting multiple-choice questions on the Kentucky Core Content Tests (KCCT) and weighting them 33 percent; weighting open-response questions 67 percent of the KCCT scores
- Giving schools incremental credit for novice and apprentice growth in Reading, Mathematics, Science and Social Studies: nonperformance is 0 points, medium novice is 13, high novice is 26, low apprentice is 40, medium apprentice is 60, high apprentice is 80, while proficient remains 100 and distinguished 140. For Writing, Arts and Humanities and Practical Living/Vocational Studies, nonperformance will be 0, novice will receive 13 points, and each apprentice will receive 60 points, while proficient and distinguished are 100 and 140, respectively.
- Reducing the testing window from three weeks to two weeks

When KIRIS was redesigned, the essential purposes remained the same. CATS is a school assessment system, and not an individual student assessment. Each student, while taking a valid test, is only addressing one sixth of the item pool that is being assessed, or in the case of arts and humanities and practical living/vocational studies, one twelfth of the item pool. The CATS is not valid for assessing whether a student has mastered the curriculum, but only whether a school has successfully taught the whole curriculum to the composite of its students.

The Transition from KIRIS to CATS

In 1999 a four year inquiry was initiated by HumRRo to study the effects of changing from the KIRIS to the CATS testing system using a purposeful sample of schools in the Commonwealth. This inquiry was to go beyond the direct uses of the test scores themselves and to focus on the following questions:

1. Are changes in a school's assessment scores reflected in changes in classroom practice?
2. Do classroom practices reflect the student learning goals stated in the Kentucky Educational Reform Act?
3. Have teachers' classroom assessment practices changed, for example, are teachers using self-designed open-response items similar in content, format, and problem-solving reasoning, to those in the KCCT for their classroom assessments?

Thirty-one schools (15 middle, 16 elementary) participated in the second phase of a four-year, four-phase, project examining the transition between the Kentucky Instructional Results Information System (KIRIS) and the Commonwealth Accountability Testing System (CATS). The first phase (Thacker, Koger, Hoffman, & Koger, 1999) included 20 schools (10 middle, 10 elementary), all of which were also included in Phase 2. Schools were selected purposefully to characterize Kentucky geographically and to include a wide range of academic performance levels.

Information was collected by observing classes, obtaining assessment materials, and interviewing teachers, principals, and district assessment coordinators. The study was qualitative and somewhat exploratory in nature, allowing researchers to investigate related, but unanticipated, topics as they were discovered in each participating school. Findings from this phase of the research show that not much has changed in response to the shift in testing and accountability systems. Schools are continuing their efforts to improve and their methods for doing so are not radically different than they were during KIRIS. Writing is still the major scholastic focus due to the portfolio component of CATS and the open-response format questions on the Kentucky Core Content Test. Teacher stress levels seem to be reduced because to the reduction in the number of required portfolio entries, however there has not been any reduction in the amount of class time spent working on portfolios. Teachers report that they are either requiring students to produce the same number of portfolio entries as was previously required under KIRIS but are requiring one less be included in the student's portfolio, or teachers and students are spending more time (editing and polishing) on the reduced entry set. Unlike KIRIS the KCCT includes multiple-choice questions in the calculation of the accountability index. These additional multiple-choice questions have not caused an appreciable reduction in the focus on open-response questions. Teachers report minor changes in instructional practices

because of the introduction of multiple-choice questions. The majority of changes in instruction include testing strategies on how to respond to multiple-choice questions, and the introduction of more multiple-choice questions in classroom tests.

Some teachers still worry about the consistency and reliability of open response scoring, student population differences, cohort effects, the breadth of the tested curriculum, test administration procedures, and other issues that might affect either their own school scores or the scores of schools that they perceive as their competition. Teachers have heard of stories of excellent students performing poorly on the test, testing violations at other schools, perceptions regarding the collective intelligence of one class versus another, and other anomalies, which give pause to the accountability system.

Thacker, *et. al.*, note that schools do pay attention to information from the Kentucky Department of Education. The Kentucky Core Content for Assessment document is clearly driving curriculum. Assessment results guide schools in developing their Comprehensive School Improvement Plans and in their efforts to improve instruction. Professional development, class schedules, programmatic changes, and resource allotment are all greatly influenced by the Kentucky Core Content Test.

Consequences: Fair to Schools

It is important that the Kentucky accountability program provide a fair educational goal for all schools. This is especially true regarding the consequences of rewards and assistance based on the assessment results. Several factors are examined to explore whether the Kentucky accountability program is fair to schools. The factors include geographical location of school, racial/ethnic composition, economic status of students, initial baseline score, school size, and grade level organization. Based on these analyses the Kentucky assessment program appears to be fair in that rewards and assistance are distributed across these dimensions without statistically significant unevenness. The exception is grade level, where proportionally more elementary schools receive rewards than do middle schools or high schools. This result seems to be explained by differences in breadth and complexity of the knowledge and skills presented at the different school levels (Elementary, Middle and High) rather than any bias.

Program-Specific School-Level Effects

Beyond being fair with regard to characteristics of student enrollment, the KCCT should not disadvantage schools participating in programmatic efforts to improve student learning. The effort in which Kentucky schools participate most widely is Title I, a federal program established to serve economically disadvantaged students by providing supplemental funding to schools, based on the poverty level of students in the district and school. Between the 1997 and 2000 school years, about 70% of the public schools in the Commonwealth participated. Kentucky is an unusually high poverty state. In 2000, 641 of 1234 (52%) school had a student poverty rate (measured by the number of students qualified for free or reduced price school lunch) of at least 50%.

Table 14-3 indicates the numbers of Title I schools participating in school-wide programs, targeted assistance, and the totals and percentages compared to all schools. Table 14-3 clearly demonstrates the decrease in use of Title I as students progress through the school system. Most of this change is the result of decreasing numbers who received free or reduced price lunches, the primary economic criteria for Title I participation. Many reasons have been proposed for this decrease in participation, increasing family wealth with the passing years, student employment, and increasing embarrassment over receiving free lunch as the students progress from the elementary grades to high school are the favorite explanations of the decrease in eligibility for Title I.

Elementary, middle, and high schools that participate in the Title I program have in the past achieved relatively greater progress toward their improvement goals than do non-Title I schools. This is a favorable finding with regard to consequential validity. However, but because of the major changes in the testing and accountability system, the trend lines between KIRIS and CATS have been severed. Thus words like 'gain,' 'growth,' 'improvement,' or 'decline' are not appropriate ways to describe the scores of 1998 and years prior as compared to those of 1999 and 2000. Hence, we are not able to describe gains for the current accountability cycle for Title I schools.

Table 14-3
A-1 School Title I Participation

	School Wide	Targeted Assistance	Total Title I Schools	Total Public Schools	Title I Percentage
GRADE 4					
1995	119	595	714	796	89.7%
1996	157	544	701	792	88.5%
1997	418	276	694	786	88.3%
1998	490	179	669	779	85.9%
1999	517	157	674	770	87.5%
2000	533	141	674	768	87.8%
GRADE 5					
1997	403	272	675	770	87.7%
1998	476	177	653	766	85.2%
1999	507	156	663	761	87.1%
2000	518	141	659	759	86.8%
GRADE 7					
1997	124	99	223	343	65.0%
1998	153	66	219	338	64.8%
1999	174	58	232	333	69.0%
2000	172	49	221	332	66.6%
GRADE 8					
1995	30	252	282	354	79.7%
1996	44	192	236	348	67.8%
1997	119	99	218	339	64.3%
1998	148	66	214	334	64.1%
1999	169	58	227	329	69.0%
2000	167	49	216	328	65.9%
GRADE 10					
1999	28	14	42	234	18.0%
2000	34	11	45	234	19.2%
GRADE 11					
1995	1	82	83	236	35.2%
1996	1	39	40	234	17.1%
1997	13	23	36	233	15.5%
1998	24	16	40	237	16.9%
1999	28	14	42	234	18.0%
2000	34	11	45	234	19.2%
ALL GRADES COMBINED[1]					
1995	129	843	972	1274	76.3%
1996	173	722	895	1274	70.3%
1997	506	372	878	1272	69.0%
1998	608	248	856	1268	67.5%
1999	651	217	868	1238	70.1%
2000	667	193	860	1234	69.7%

[1]The total is smaller than the sum of the grade levels because of the P-8. P-12, and 7-12 schools and school that have multiple charted grades.

Conclusion

The results described in this chapter indicate that extensive resources, large amounts of time, and significant effort have been devoted to validity. The goal is to establish, maintain and improve the KCCT to support and encourage efforts to improve the educational achievement of each child in Kentucky.